EQUIVALENT DESIGN PROBLEMS, AN EXPERIMENTAL STUDY

A Thesis
Presented to
The Academic Faculty

by

Bryan Levy

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
G.W. Woodruff School of Mechanical Engineering

Georgia Institute of Technology
August 2017

EQUIVALENT DESIGN PROBLEMS, AN EXPERIMENTAL STUDY

Approved by:

Dr. Julie Linsey, Advisor
School of Mechanical Engineering
Georgia Institute of Technology

Dr. Raghuram Pucha
School of Mechanical Engineering
Georgia Institute of Technology

Dr. Chris Paredis
School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: May 2nd , 2017

ACKNOWLEDGEMENTS

I would first like to thank my advisor Dr. Julie Linsey for allowing me to study under her guidance. Through her support, I was able to gain immeasurable insight into a new area of research and grow as both a researcher and as a person. I would also like to thank Dr. Raghuram Pucha and Dr. Chris Paredis for reviewing my thesis and agreeing to be on my committee.

Additionally, I would like to thank all of my lab mates who have helped me along the way. I would like to thank Ricardo Morocz for all of his help and friendship in our initial research together, and Megan Tomko for her advice and organizational skills. I would especially like to thank Ethan Hilton for his support in the lab and on this project in particular. Without his help in establishing inter-rater reliability, I would not have been able to complete my studies.

I would like to acknowledge that the support for this work was provided by the National Science Foundation Award No. DUE-1432107. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

I would like to thank my parents, Mary and Gary Levy for their unending support of my academic and personal growth. Finally, I would like to thank Katie Scott for everything she has done for me. It is with her support and help that I was able to achieve this accomplishment and am truly grateful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

A current standard for testing advances in engineering design is to present participants with a design problem and evaluate their performance across metrics. This approach, however, cannot be used in repeated measures testing, as a design problem cannot be given repeatedly to a participant without biasing results, therefore necessitating equivalent design problems. This study provides a foundation for creating these equivalent design problems by investigating four design problems for equivalency using between-subjects and within-subjects testing. These problems have been modified for greater similarity and include peanut, corn, alarm, and coconut design problems. All design problems were given during between-subjects testing and the peanut and corn problems were also tested within-subjects. The within-subjects analysis revealed correlations on three of five metrics tested, including quantity which is the most frequent metric in the field [1, 2]. This indicates that the problems used are close to equivalent and may be used in repeated measures studies for these metrics. Additionally, the between-subjects analysis revealed that the design problems do not show group mean equivalency, meaning between-subjects analysis is insufficient in assessing equivalency and within-subjects analysis should be the standard in future studies of design problem equivalency. Three problem characteristics [3] were also investigated in the between-subjects analysis for their impact on metric scores. While trends emerged, such as higher difficulty leads to fewer and lower quality solutions, more research needs to be done to understand how manipulating these problem characteristics and couplings of characteristics can be used to make equivalent design problems.

CHAPTER 1. INTRODUCTION

1.1 Motivation for Research

In order to enhance the creative process, researchers are continuously developing new methods to guide the design process. Currently, advances are being made to characterize why designers react a certain way to a given design problem [3-8], and how design changes with different levels of expertise [9, 10]. Additional research is underway to track the changes in design thinking, more specifically design creativity, over the span of an undergraduate curriculum [11]. For evaluating design, a current standard in engineering design presents a design problem to participants and asks them to generate solutions [12-18]. Typically, the design problems are sculpted for a particular study or recycled from previous experiments [3, 19]. A multitude of these problems having been crafted over the years – each with its own individual purpose [19]. The solutions generated from these problems are generally scored across the metrics of quantity, quality, novelty and variety originally put forth by Shah et al. [20, 21].

For studies of this nature, researchers must choose between a between-subjects and a within-subjects approach. A between-subjects approach is typically selected for this research [4, 22-26] as it is easy to implement and allows researchers to use a single design problem, eliminating the outside variable of inconsistencies in design problems. However, this method decreases sample size in each group, assuming a finite number of available participants, and presents the possible bias of non-representative sampling within the population. A within-subjects approach may therefore be useful in a number of situations, namely to increase the precision of the measurement [27, 28], or to eliminate

individuals differences. Bias due to individual differences is exemplified where longitudinal studies are utilized to minimize the effects of different populations or to track particular populations. While useful in reducing population composition variations and increasing statistical power, within-subjects design presents its own drawbacks. Within-subjects design may bias results as participants may become sensitive to the purpose of the experiment and react differently or may carry-over previous experience gained between testing sessions [28].

In situations where within-subjects design is beneficial, researchers must try to reduce some of the potentially biasing factors. A potential way to mitigate the carry-over effect of experience with a particular design problem would be the use of a new design problem for each new test session. Currently, this is not possible due to empirically demonstrated similar problems not being available. Within-subject experiments therefore would greatly benefit from equivalent design problems.

1.2 Research Scope

This paper is a step in developing design problems that are equivalent, so that they may be used interchangeably in within-subjects research. In this work, four design problems (peanut, corn, alarm, and coconut) are evaluated for problem equivalency, where equivalency is characterized based on the metric scores of the solutions for each design problem. Two approaches for analysis are executed: a large-scale between-subjects analysis and a smaller within-subjects analysis. The between-subjects analysis investigates the four design problems in terms of group mean similarity. A corresponding survey of these participants gathered data on certain characteristics of the design

problems, suggested in earlier work [3], to observe their effects on metric outcome. Furthermore, the within-subjects analysis investigates the correlations between two of the problems: peanut and corn. Additionally, this study looked into the possibility of utilizing between-subjects design as a first step towards design problem equivalency. While within-subjects design is presumed as the preferred method for testing equivalency, it may be found that between-subjects testing can serve as a preliminary screening method for finding equivalent design method. As both methodologies are used in this research, the research presented will be able to assess the validity of using between-subjects as a screening method.

1.3 Thesis Organization

The remainder of this thesis is outlined in the following manner. Chapter 2 presents a review of background information on design problems as well as current research on variables affecting the results on creativity as measured according to ideation metrics. Chapter 2 additionally provides background information on the techniques used in this paper to assess the equivalency of the design problems. Chapter 3 describes the experimental procedure used in all parts of the study and a more in depth description of the ideation metrics used in the study. Chapter 4 presents the results found over the course of the study as well as a description of their implications. Chapter 5 presents a conclusion of the results of the study and Chapter 6 outlines steps to be taken in future works building upon this study.

CHAPTER 2. BACKGROUND

A key part of the graduating engineers' repertoire is the ability to think creatively in solving problems presented to them. This link between engineering and creativity has been present since the late 50's where it was introduced in work by Sprecher [29] where he first investigated how engineers assess creativity. More recently, work by Cropley, [30] reasserts the importance of creativity to engineers, outlines the creative process, and points out difficulties in translating creativity into engineering education. To this end, it is important that as educators we do our best to foster this skill in students. However, a difficulty arises from the lack of an ability to assess creativity and changes in creativity in a consistent manner. Despite this difficulty, an area of design that has shown potential as an indicator of creativity in engineering design is the conceptual design phase. The conceptual design phase is a stage of design where designers or engineers develop, assess, and select a design to solve a problem [31]. The conceptual design phase is currently used for assessing creativity in engineering design by presenting students with a design problem and grading their responses across several metrics that relate back to creativity. This technique has been in place with some small variations since the early 2000's [20].

In a recent study by Kumar and Mocko, the reuse of design problems, or lack thereof was made apparent. It was found that over the last 15 years, over 46 different design problems have been used in less than 46 research papers [19]. This lack of design problem reuse is an obvious concern for the design community because it makes it difficult to make reliable advancements and comparisons across research groups. As

different design problems are shown to affect metric scores in different ways [3, 19], the use of different design problems in the field eliminates the possibility of direct comparison of the results across studies without rerunning the studies using the same problems. Additionally, a recent psychology study looking at the repeatability of publications in psychology found a surprisingly low number of studies, approximately 40 percent, met their guidelines of reproducibility [32]. The guidelines in use included obtaining the same statistically significant effect while maintaining the same effect size as the original study. When the guidelines for the study were extended to include similar, but not statistically similar results, that number of reproducible studies is higher, close to 65 percent [32]. Although the repeatability study is focused on another field, it still raises concern about the importance of repeatable results. The research presented in this paper serve to increase the repeatability of studies within engineering design by presenting a single set of design problems to be used across design research.

2.1 Design Problems

As a key part of the evaluation technique, the design problem is a subject of study for design researchers. Design research has postulated that the structure of the design problem plays a significant role in the corresponding solutions generated. Originally, the design problem was said to define the problem space, which in turn defines the possible solutions available for the participant to find [33]. This was later refined to say that the "ill structured" nature of design problems forces the participant to solve the design problem in smaller temporary design spaces that are part of the larger solution space [34]. Alternatively, a theory was postulated by Schön that states that design is not a straightforward process but rather one that requires reflection [35]. According to this

theory, the structure of the design problem does not directly influence the results but rather the framework through which the designer uses. Similarly, the design process has been described as an adaptive process where the problem and solution space simultaneously co-evolve in an iterative process [36, 37]. This theory is further explored by Dorst where he argues that the design process can be better approximated by the relative expertise of the designer as opposed to the structure of the design problem itself [38]. In other words, the design process used by the participant is predetermined from the experience level of the designer and is not primarily driven by the design problem.

Further studies have been made looking into design problems and the role of their structure and underlying nature. Studies have shown that a design solution may be forecast to an extent based on the complexity of the problem in terms of size and coupling [7, 39-41]. A system was then outlined to help researchers understand how to track this complexity throughout the design process [42]. Other research has suggested that the semantics and writing style of the problem can have a non-negligible influence in the manner in which the problem is solved [19]. The end goal of these studies is to understand to what extent the design problem and the nuances of its language affect the solutions generated by the participants. To this end, Durand et. al. have put forth a list of characteristics of design problems believed to influence the design results, such as size, connectedness, and familiarity with the problem and solution spaces [3]. This list of characteristics and their associated literature as well as additional characteristics found in literature can be seen here:

- Size of the problem in terms of functional units or components [7, 39, 40, 42]

- Connectedness of the problem in terms of coupling between functional requirements or constraints [7, 39, 40, 42]
- Size (number of variables) of the potential solution space, and the degree to which they are constrained [7, 39, 40, 42]
- The degree to which existing solutions will cause fixation [5]
- Participants' familiarity with the design problem, and the underlying principles inherent in the problem [3, 43]
- Participants' familiarity with the existing design solutions, and the underlying principles required to generate a solution [3, 43]
- Assumed constraints due to known solutions, culture or other factors [3]
- The effort required to solve the problem, in terms of the degree to which the problem is technically challenging [3]
- The domain of the design problem, and the degree to which ex-domain analogous solutions are easily retrieved [44-47]
- The semantic presentation of the design problem [19]

In the field of design, a set of equivalent design problems is desired for use in any form of repeated measures analysis utilizing idea generation. While significant research has been done on the structure and wording of design problems [3, 5, 7, 19, 39-42], little research has been done on the actual equivalency of the design problems with respect to the solutions generated [3]. Several studies have claimed to use design problems that they assert as equivalent [13, 48, 49], but without further testing this claim is still undetermined. To eliminate the possibility of design problem inequalities biasing repeated measures testing, the research in this study investigates the equivalency of four

design problems used in the field for problem equivalency. The information gained through this study will provide knowledge on the equivalency of these design problems and will improve upon the framework used for evaluating these problems. The results will provide a platform for future research in testing the equivalency of design problems.

2.2 Test Setup

When thinking about repeatability and consistency across studies, it is important for researchers to try to use a standard testing procedure. A standard testing procedure reduces the effect of added variables while ensuring easier reproduction of results. The most common procedure to date is to give participants an open-ended design problem and a set amount of time to generate solutions during a single session [21]. However, the amount of time given to participants varies between studies as an ideal time is uncertain. Liikkanen and colleagues have found that on average the test time given ranges from 20-60 minutes [50]. This time range may have been used as researchers feel that fatigue diminishes results after this point, or perhaps a more likely explanation is that it is easier to secure participants (oftentimes students) for shorter time durations.

Whatever the reason, the amount of time given can have an effect on the results obtained by the study. It was found across multiple studies that the number of ideas generated by participants' decreases over time with the largest number of ideas generated at the beginning of the testing procedure [50-54]. More important perhaps than just the number of solutions generated is the content of the solutions. It was theorized by Guilford [55] that at the beginning of concept generation participants will produce, common well-known solutions, and over time the solutions generated will be more novel and unique. In

studies that looked at ideas generated over time, it was found that participants that were given two hours as opposed to one hour generated more solutions and a larger variety of solutions [23, 56]. The solutions generated after the first hour did not have any greater novelty or quality scores and were still functional solutions to the design problem.

An incubation period has been suggested in order to help alleviate fatigue in participants while extending their generation time by splitting generation into multiple sessions[57]. This incubation time has been tested and it was found to help alleviate the phenomenon of design fixation, discussed in Section 2.3, in participants [23, 57, 58]. However, when compared with a single longer idea generation session, participants produced more novel and functional ideas during the single session [23]. This suggests that an incubation period is more useful for problems known to exhibit higher degrees of fixation. Variance in test times and test format can therefore have a significant impact on what solutions are generated. Extended test times may be beneficial to allow students to more completely explore the solution space, while an incubation period can help alleviate design fixation. However, extending the test time may limit the ability to recruit participants and may fatigue some participants while an incubation period can similarly be harmful to recruiting. The design community must therefore reach a consensus that balances the benefits of extended test times with the resources most readily available, oftentimes one-hour periods corresponding with the length of classes. The test setup can be varied to some extent, however, depending on the goals of the researchers while maintaining the same general guidelines. For example, the research team may look at individual performances or utilize group design techniques such as 6-3-5, group brainstorming, etc. by only manipulating participant interactions.

2.3 Design Fixation

Another key part of the design problem is whether an example is given to the participant as part of the design prompt. To this end, several studies have been conducted on presenting examples to participants as part of the design problem. Jansson and Smith first documented that exposure of participants to example solutions leads to design fixation [5]. Design fixation occurs when participants use parts of an example in their answer to a greater degree than they would without the example. This process can be either beneficial to the designer or harmful depending on what is being repeated. Further studies have looked deeper into the topic, in order to understand the nuances causing fixation. The quality and type of example given have been investigated for their impact on the process. Purcell and Gero originally suggested the effects of the quality and types of examples given [59, 60]. They found that exposure to more typical solutions caused greater fixation, however the opposite does not seem to be true where exposure to novel ideas leads to the development of more novel solutions [61]. The quality of the example given can play a factor in the end result as seen in studies of poor versus good examples, where it was found that poor examples lead to greater fixation [61, 62]. Possibly more important still to designers is what is being copied from the examples. It was found that poor examples could cause designers to copy over some of the features that made the design poor without their knowledge [25, 56, 63]. The effects of multiple examples as opposed to a single example have been explored [6, 51, 61] where there appeared to be no significant difference in number of examples given. It is important to note that results from social psychology experiments suggest that there could have been [64, 65]. The type of example can also factor into the degree of fixation by the participant, as found by

Wilson et al. [8], where they found that introducing a surface dissimilar example led to lower novelty but higher variety solutions than a surface similar example.

An additional topic of research for those investigating fixation is the manner in which the example is presented. Typically, examples are presented to participants through text, sketches, or photos. Textural versus pictorial based examples were investigated by McKoy et al.[66] where it was found that when presented with a pictorial example, participants scored higher in quality as well as novelty metrics than those presented with an example written in text. Additionally the study found that sketch based solutions were easier to comprehend and generally scored higher than strictly text based solutions. One study looked at the differences when presenting examples as photographs versus as line drawn representations [67] and found that both methods led to fixation with no significant differences across quantity, quality, and originality metrics. Although no statistical difference was found, data suggests that the participants exposed to the line drawing may produce more novel ideas than the photographic examples. A similar study conducted using a good example (presented as a sketch, a picture, and a CAD image) showed similar results where fixation was found in similar amounts across all three media [15]. In this case, the fixation lead to higher quality results as participants more easily repeated successful features. In another study, Goldschmidt and Smolkov addresses how these fixations can actually be beneficial on the quality of results but the effects of visual examples vary depending on the design problem [68]. An additional study looked at exposing participants to a function tree, a typical tool in decomposing a design problem, as an example media and found that using a function tree does not lead

to fixation and pairing this with a sketch example can strengthen feature fixation while reducing idea fixation [4].

2.4 Ideation Metrics

An important step in testing participants through idea generation is measuring and assessing performance of the participants with respect to the design problem. Shah et al. put forth four metrics to use in evaluating solution performance: quality, quantity, novelty and variety [20, 21]. These original four metrics work by decomposing the design into its base functions and then grading each solution by how it addresses the functions and taking the weighted sum to get each metric. Since then, various research teams have modified the metrics to meet their needs. Linsey et al. have adapted the metrics by removing the weighted sum process from the metrics to avoid bias in weights [69, 70]. Additionally, the designs are no longer broken down by functions but solutions can be broken down according to functional basis [71] in order to facilitate metric scoring. Looking at solutions at the conceptual versus the feature-based level has been studied before, and it was found that scales with fewer increments are more repeatable but can lose precision [72, 73].

Individual metrics have also been reworked by different research teams in order to improve upon original shortcomings. Nelson et al. proposed reworking the variety and the quantity metric into a single metric that evaluates the exploration of the design space in order to eliminate redundancies [74]. Several ideas have been put forth for changes to the novelty metric. A weighted sum approach has been suggested [75] that breaks features down by the type of design (adaptive, novel, redesign) and assigns novel design

features greater value. An alternative approach to novelty is the originality metric formed by Charyton et al. [76, 77] that rates concepts or features on an eleven point scale from dull to genius. The originality and Shah's novelty metrics have been compared and further studied to understand the benefits of each [26, 72]. The novelty metric suffers a setback as it compares novelty to the solutions generated during that session. This means that features that are novel to the market are not scored as novel if that feature was frequent during the session. This problem is especially noticeable during group brainstorming techniques as a single feature may be incorporated into multiple solutions by a single author. The originality metric, however, suffers due to its subjectivity, as the score is largely based on comparisons made by the grader to the existing market. Further examinations into the novelty metric assert that it is an area requiring greater attention due to the weaknesses present in its different alternatives [78]. A more in depth description of the metrics used in this study can be found later in the paper in section 3.4.

2.5 Testing for Equivalency

The objective of this study is to observe the equivalency of four design problems for use in design research. This statement of intent can be reworded to better conform to traditional validation methods by stating the objective of the study is to assess the reliability of a creativity test. This creativity test, as described in section 2.2, presents a participant with a design problem and grades solutions generated across ideation metrics, as described in sections 2.4 and 3.4. Reliability can then be evaluated in two ways: internal-consistency reliability and test-retest reliability [79]. Internal-consistency reliability measures how well a test measures a certain variable. This is done by gauging the correlation of different questions on the same test designed to assess a particular

variable. While this internal-consistency reliability is an important characteristic of a test, it becomes difficult to measure as the test consists of a single open-ended question. Therefore, this form of reliability is not under investigation in this study.

Test-retest reliability, as the name implies, looks at how the scores of a participant are related when tested on separate occasions (assuming the participant did not change) [79]. For the creativity test in use for this study, it is unrealistic to give the same design problem, and hence the same test, in short succession to prove the test-retest reliability metric. However, due to the structure of the testing procedure, each design problem can be viewed as a parallel form of the same test. In this way, test-retest reliability can be estimated for parallel forms using the same procedure which looks at correlation values using within-subjects design [79].

An assumption corresponding with test-retest reliability is the presence of a true score. True score theory states that for a particular trait, an individual has a true score and tests are designed in order to ascertain this true score. For this study, this would translate to individuals having a specific score for creativity that the creativity tests try to measure. Test-retest reliability is then a measure of how reliable the test or parallel forms of the test measure this true score. As with other statistics, sample size is an additional concern for test-retest reliability. Previous literature [80] has voiced concern that sample sizes close to 40 or 50 in each population may be insufficient to yield a stable estimate. As, the sample sizes in this study, fall within this range, it is important to note that further expansion and retest with yet greater numbers may be needed to confirm the results for the field.

In this study, test-retest reliability is measured for parallel forms by correlating individuals' scores for the ideation metrics. This is accomplished through within-subjects experimentation. Between-subjects experimentation is additionally used in the study to understand the relationships the design problems share. This will help future iterations of the design problems reach equivalency by controlling the relationships. The between-subjects experiment also serves to determine if it is plausible to estimate test-retest reliability through between-subjects experiments. If shown as plausible, this would allow researchers to utilize an easier and cheaper methodology to determine if parallel forms were sufficiently reliable for use.

2.6 Equivalency Tests

In this study, equivalency of design problems is assessed utilizing statistical testing techniques to compare the metric scores of design problems with respect to each other. As discussed in section 2.5, equivalency is assessed according to correlations of individuals' scores in within-subjects testing. Additionally, tests for equivalency according to between-subjects data are used in the case that this may serve as an estimator of true test-retest reliability. To do this, a Mann-Whitney test for equivalence is utilized accounting for nonparametric tendencies, such as non-normality. This technique was previously developed [81] for use in bioequivalence and a summary of this technique is presented in this paper. Additionally, the concept of confidence interval inclusion is demonstrated for use in parametric data. Derivations of these tests as well as other equivalency testing methods can be found in previous works by Wellek [81, 82].

2.6.1 Underlying Theory

Modern statistical techniques for comparing two sample means test the data against two hypotheses (null vs. alternative) in order to assert the alternative hypothesis with a degree of confidence (1-α). The null hypothesis being that the means ($\mu_{1,2}$) are equivalent H_0: $\mu_1 = \mu_2$ and the alternative hypothesis stating the opposite $H_{1:}$ $\mu_1 \neq \mu_2$. In other words, the statistic tests whether there is enough evidence to show statistical difference and accept the alternative hypothesis. The use of this statistical technique for equivalency can therefore be seen as flawed, as it cannot assert equivalence but rather a lack of sufficient evidence to prove a difference. To overcome this shortcoming, the hypothesis must be rewritten in order to test for equivalency as the alternative hypothesis. It is therefore necessary to define equivalency, which can be done by establishing an interval about the mean difference (μ_2 - μ_1 = θ) in which two samples are considered equivalent. This equivalence interval can be defined by a shift from the mean difference by a prescribed amount ε and should be determined a priori by the researchers. A table of suggested equivalence limits for different types of equivalency tests is suggested in literature [82] with a value of 0.2 selected for this study, corresponding to equivalency within 20%. It is possible to define an asymmetrical equivalence interval if desired, but this study utilizes a symmetric interval. With this definition of equivalency, the null and alternative hypothesis can be rewritten as seen in Equation (1) and Equation (2), respectively.

$$H_0: \theta \leq \theta_0 - \varepsilon \ \ or \ \ \theta \geq \theta_0 + \varepsilon \tag{1}$$

$$H_1: \theta_0 - \varepsilon < \theta < \theta_0 + \varepsilon \tag{2}$$

In this study, θ_0 was chosen to be equal to zero implying a mean difference of zero is desired. With this new definition of the null and alternative hypothesis, the data can be appropriately tested for equivalence. For parametric data, this is typically done using two one-sided t-tests that test individually for each null hypothesis presented in Equation (1) in a procedure commonly referred to as TOST. The data can then be stated to fall within the interval with a 1-α confidence level. This procedure can be shown mathematically to be the same as a process called confidence interval inclusion.

Confidence interval inclusion works by first establishing an equivalence interval a priori that if it contained the mean difference the user would deem the two samples equivalent. For example, one could say within 2 millimeters in a manufacturing process or within 20% of a medical effect. A 1-2α confidence interval is calculated about the null hypothesis of equal means as done in typical statistical methods. If the confidence interval is included in the equivalence interval then the data is said to be equivalent. However, the data does not always meet the assumptions needed for parametric testing and these methods cannot always be used. In these instances, it is necessary to use nonparametric statistical techniques. A similar approach to the TOST method was developed using the Mann-Whitney nonparametric test. However, this method can be restrictive as it only looks into the shift in mean values and not distributions [81]. To overcome this and provide a more complete nonparametric equivalence test Wellek put forth the Mann-Whitney test for equivalence [81].

2.6.2 Mann-Whitney Test for Equivalence

To implement the Mann-Whitney test for equivalence, one must first accept the set of null hypothesis and alternative hypothesis put forth in the previous section. These hypotheses are first adapted to fit the framework of the Mann-Whitney test statistic as follows.

$$H_0: \pi_+ \leq \frac{1}{2} - \varepsilon' \ \ or \ \ \pi_+ \geq \frac{1}{2} + \varepsilon' \tag{3}$$

$$H_1: \frac{1}{2} - \varepsilon' < \pi_+ < \frac{1}{2} + \varepsilon' \tag{4}$$

Equation (3) represents the new null hypothesis and Equation (4) represents the new alternative hypothesis where π_+ represents probability of an observation from the first population that exceeds a sample from the second population and ε' represents a shifted equivalence limit. With the hypotheses established, the test is conducted by calculating the U-statistic estimator W_+ of π_+, as well as the standard deviation of W_+. Due to the asymptotic normality of the Mann-Whitney statistic, the hypothesis test now reduces to Equation (5).

$$\frac{\left| W_+ - \frac{1}{2} - \frac{(\varepsilon'_2 - \varepsilon'_1)}{2} \right|}{\sigma[W_+]} < C_{MW}(\alpha; \varepsilon'_1, \varepsilon'_2) \tag{5}$$

Where C_{MW} is a critical value calculated based on a Chi squared distribution and the desired significance of the test as well as the equivalence limits as can be seen in Equation (6).

$$C_{MW}(\alpha; \varepsilon'_1, \varepsilon'_2) = \left\{ \begin{matrix} 100\alpha - percentage\ point\ of\ the\ \chi^2 - \\ distribution\ with\ df = 1\ and\ \lambda_{nc}^2 = \dfrac{(\varepsilon'_1 + \varepsilon'_2)^2}{4\sigma^2[W_+]} \end{matrix} \right\}^{1/2} \quad (6)$$

In this way, if Equation (5) is true then the null hypothesis of nonequivalence should be rejected and the alternative hypothesis of equivalence accepted. This test looks at the distribution of the data as well as the raw difference in means and as such asserts a more complete and restrictive determination of equivalency. In this study, equivalence was calculated in this manner utilizing a program called *mawi.R*, written for the statistics software R that is supplied by Wellek [81, 82]. The inputs for the program are the data, the desired significance α, the equivalence limits ε'_1 and ε'_2, and the sample sizes for the two groups. The program outputs the U-estimator W_+, the standard deviation $\sigma[W_+]$, the critical value C_{MW}, and the decision of whether to reject the null hypothesis based the calculations seen in Equation (5).

CHAPTER 3. EXPERIMENTAL PROCEDURE

3.1 Design Problems

As a starting point, four design problems were initially chosen from a subset of design problems in literature used by various research teams [19]. These design problems will be referred to throughout the paper as Peanut, Corn, Alarm, and Coconut due to the subject matter of the design problems and can be seen in full in Appendix A. The Peanut problem asks the participants to design a low-cost machine to shell peanuts in regions where resources are limited [3, 4, 10, 15, 23, 25, 62, 63, 83-87]. The corn problem asks participants to create a device to facilitate shucking corn [3, 87-91]. The alarm problem asks participants to design a portable alarm clock that can awaken its users without disturbing others [3, 26, 72, 87, 89, 90, 92]. The coconut problem asks participants to design a system to retrieve coconuts from tall coconut trees in a region where resources are limited [87, 93]. The design problems that were selected were similar in the technical requirements asked of the designer, and three of the four problems (Peanut, Alarm, and Coconut) were similar in their domain, nature.

The design problems were altered from their original forms in hopes of reducing variances and increasing consistency in scoring across the design problems. The author rewrote the problems in order to have the same writing style throughout. Additionally, the problems were presented in the same way; a paragraph describing background information followed by a design statement, a picture to help familiarize the designer with the problem, and a list of five customer requirements to be meet for each design. This problem setup was adapted to match the format of the problem most frequently used

in past research on idea generation, the Peanut problem. The Peanut problem was selected to be the baseline of comparison for this study due to its extensive use in the literature. The rewritten design problems can be found in Appendix B.

With this initial set of design problems selected and prepared, the problems were tested for equivalency. This was done in the form of a large-scale between-subjects investigation coupled with a smaller scale within-subjects investigation. The between-subjects testing was conducted in order to get a more complete understanding of all the design problems in terms of metric scores and problem-designer characteristics such as perceived difficulty and familiarity. The within-subjects testing was conducted in order to obtain richer, higher validity data on problem comparisons. Both between-subjects and within-subjects testing was conducted using one-hour idea generation sessions. Greater detail of the test setup is available in the following sections.

3.2 Between-Subjects Testing

For the between-subjects portion of this study, an idea generation activity was conducted with students in a freshmen level engineering design course. Students in this course include freshmen and sophomore level Mechanical and Aerospace engineering students and occasionally junior and senior Aerospace students. During a scheduled three-hour lab session, the students were given the opportunity to participate in an hour-long individual idea generation activity for extra credit.

Those that chose to participate (approximately 87% of the class) were given of the four design problems at random, alternating through the problems to try to maintain an equal distribution. The students were given 2 minutes to read the design prompt. At the

end of these 2 minutes, any questions were fielded and the students were then given 48 minutes to generate as many solutions to their design prompt as possible in the form of a sketch and accompanying description. During this time, the students were not allowed to interact with other students or use outside devices (music players, cell phones, computers, etc.). Following the idea generation session, the participants were given a short survey on perceived difficulty and problem familiarity with the question format taken from literature [94]. This survey can be seen in Appendix C.

Over the course of the between-subjects testing, 198 students from five sections took part in the idea generation activity. Of those students, 190 students completed the activity while eight students did not follow the correct procedures, such as talking, listening to music, or leaving early, and their results were excluded.

3.3 Within Subjects Testing

The within-subjects portion of the survey was conducted using a different sample of participants. This part of the experiment was conducted using students in the mechanical engineering capstone/senior design course. During one week of the semester, students were given the opportunity to participate in an idea generation activity over the course of two lectures for extra credit and/or cash compensation. Those students that chose to participate took part in two 50-minute ideation activities of the same format as the between-subjects study. However, for the within-subjects participants, the design problem distribution was not random. All participants were given the peanut design problem, as this was the baseline design problem, during the first session and the corn design problem two days later during the second session.

As there were only enough resources to directly compare two design problems via within-subjects analysis, the peanut and corn design problems were selected for testing. The peanut problem was selected as the baseline and the corn problem was chosen as it was hypothesized to have the closest initial similarities to the peanut problem. These similarities includes similar technical difficulty (approximated by the author), size of the problem in terms of functional units [3], and a similar problem domain (removing the outer casing of an object found in nature).

Over the course of the within-subjects testing, 83 students took part in the idea generation activity. Of those, 40 students completed both ideation sessions and one student did not follow all the correct guidelines during testing, such as talking during the test.

3.4 Ideation Metrics

Upon completion of the idea generation activity, the students' documents were anonymized and the submissions graded according to metrics originally developed by Shah [20, 21] and later improved by Linsey, et al., [69, 70, 83]. The metrics used in this study were Quantity, Quality, Novelty, Variety, Number of Solutions, and a new supplementary metric of Completeness.

3.4.1 Quantity

The quantity metric used in this study is adapted from previous literature [4, 15, 20, 21, 95] and measures the quantity of unique ideas presented by the participant. It is important to define what constitutes as a unique idea for this study and what constitutes

as a solution. A single idea is defined by something that solves one or more of the functions of the design as defined in previous literature [83, 96]. A solution is defined as a collection of one or more of these ideas presented by the participant as a single design. The quantity metric is calculated by counting the number of non-redundant ideas present across all solutions of a participant.

3.4.2 Quality

The quality metric assesses the feasibility of a solution variant and how it meets the customer needs [21]. In this study the metric is graded according to a 3-point scale developed by Linsey et. al. [96]. A score of zero is assigned to the solution if it is deemed not feasible from a technical standpoint or if its implementation would not satisfy the customer needs put forth by the problem. A score of one is designated to a solution that is feasible and partially fulfills the customer requirements and a score of two is awarded to a solution that is feasible and satisfies most or all of the customer needs of the problem. The quality metric reported in this study averages the quality scores of all solutions produced by the participant. In this study, the quality metric is calculated with the aid of the Completeness metric.

3.4.3 Completeness

In an attempt to alleviate the discrepancies found in inner-rater reliability with the quality metric, a six-point completeness metric was developed. The completeness metric works as a supplementary metric that takes advantage of the fact that all design problems used have the same number of customer needs. The metric applies a binary score to each solution on each customer need established in the problem description. Additionally, if

the design solution is not technically feasible or does not solve the fundamental problem, the solution receives a completeness score of zero. The score is then summed across all customer needs to achieve a completeness score. This can be seen below in equation (7) and graphically for an example set in Table 1

$$Completeness = Technical\ Feasibility * \sum Customer\ Needs\ achieved \qquad (7)$$

Table 1. Completeness Metric Example

Participant	**A.1**	**A.2**	**B.1**
Is the solution technically feasible?	1	1	0
Does the solution meet or exceed the following customer needs?			
Customer Need 1	1	0	0
Customer Need 2	1	1	0
Customer Need 3	0	1	1
Customer Need 4	0	1	1
Customer Need 5	1	1	1
Customer Need Total	3	4	3
Completeness Score	1 * 3 = 3	1 * 4 = 4	0 * 3 = 0

This can supplement the quality metric by supplying guidelines for meeting most of the customer needs (the difference between a score of 1 and a score of 2). For this study the quality metric is rated according to a three point system (0, 1, 2) and all problems are assigned to have five customer needs, so the quality metric score can be found by translating completeness scores into quality metrics. A completeness score of zero or one translates into a quality score of zero, a completeness score of two or three translates into a quality score of one, and a completeness score greater than three translates into a quality score of two.

3.4.4 Novelty

The novelty metric measures how unique a solution is in comparison to all solutions generated during that idea generation session [21, 74]. This score is calculated using a bin system where the solutions are first sorted into one or more bins according to the procedures previously set forth in literature [83, 96]. Each design problem uses its own bin list that can be compiled from previous experiments or created new if no empirical data is available. In this study, existing bin lists were retrieved for each of the design problems and expanded for the corn, alarm, and coconut problems as new solution categories were presented in the solutions generated during this study. The final bin lists can be seen in Appendix D. The novelty metric is calculated using equation (8) that calculates the novelty of each bin and assigns each idea within a solution the novelty score of its associated bin.

$$Novelty = 1 - \frac{Total\ \#\ of\ ideas\ in\ bin}{Total\ \#\ of\ ideas\ binned} \tag{8}$$

The final novelty score for a particular solution can then be found by averaging the novelty values for all binned ideas within that solution. Novelty scores can be expressed as the average novelty of the participant by averaging the novelty values of all the participant's ideas or as the maximum novelty of the participant, which is the highest novelty score that a participant receives for a single solution. For this study, the average novelty of the student will be used.

3.4.5 Variety

The variety metric measures how much of the solution space is explored by each participant during the ideation session [21, 74]. This metric utilizes the same bin list as the novelty metric [83, 96]. To calculate the variety score, equation (9) is used which looks at the number of different bins used by a participant in comparison to the total number of bins created for that design problem.

$$Variety = \frac{Number\ of\ bins\ a\ participant\ used}{Total\ number\ of\ bins} \tag{9}$$

As variety measures the exploration of the solution space over the entire session, the metric is a single calculation for each participant using all of their solutions.

3.4.6 Number of Solutions

As this study looks at the equivalency of design problems, the number of solutions generated for each problem is investigated [3]. For this study, a solution is defined as the collection of one or more ideas presented by the participant as a single design. The number of solutions metric differs from the quantity metric in that the quantity metric counts unique ideas across all solutions of a participant, whereas the number of solutions metric counts all solutions presented by the participant.

3.4.7 Inter-rater Agreement

The reliability of the metric grading was assessed for the data using the following grading scheme. All of the data was graded by the author originally and was checked for

repeatability by having a single second grader rate the solutions for 10 participants chosen at random from each group representing approximately 20-25% of the data. The scores were compared using a Pearson's correlation for Quantity, Novelty, and Variety, and compared using Cohen's Kappa for Quality. The results of the inner rater agreement can be seen in Table 2.

Table 2. Inner-Rater Agreement

	Between-Subjects								Within-Subjects			
	Peanut	n	Corn	n	Alarm	n	Coconut	n	Peanut	n	Corn	n
Quantity	0.75	10	0.82	10	.93	10	.94	10	0.82	10	0.84	10
Quality	0.18	10	0.48	48	.62	10	0.22	10	0.44	10	0.23	10
Novelty	0.81	10	0.74	10	.91	10	.65	10	0.78	10	0.92	10
Variety	0.58	10	0.64	10	.99	10	.80	10	0.96	10	0.62	10
Percent of Sample	20.00		20.83/100		21.74		21.74		25.64		25.64	

As can be seen from the table the inner-rater agreement varies between problems but for the most part, there is an acceptable to good agreement between raters. The quality metric has the largest variability between raters. This can be caused from expertise differences in assessing quality and largely speaks to the subjectivity of the current quality metric. Due to initial low inter-rater reliability with the quality metric of the corn problem in the between-subjects experiment, the second grader graded the rest of the participants' solutions for this metric. This caused an increase in the inter-rater reliability, and it is the belief of the author that increasing the percent of the sample checked by the second grader would increase reliability across all metrics. The original inter-rater reliability can be seen in Appendix E.

In processing the inter-rater reliability, the author noticed a trend needing further study where grader familiarity seemed to play a role in agreement. In problems where both graders had similar knowledge of known solutions or similar practice grading a particular problem, the agreement in more subjective scoring metrics, such as quality, were improved. This trend, if shown in other studies, would mean that comparing scores across research groups and even within research groups would become very difficult without a single grader. This finding reaffirms that improvements are still needed in order to remove subjectivity from the ideation metrics.

CHAPTER 4. RESULTS

4.1 Between-Subjects Testing

Once the data was collected, the students' ideation documents were de-identified and graded according to the metrics described in section 3.4. The sample sizes, section and design problem can be seen in Table 3. The data was first analyzed to determine the effects of outside variables, such as class section, by testing for statistical differences in the data and none were observed. Once this was completed, the data for each design problem was combined and analyzed. Differences between the design problems were checked using t-tests when the data was normal and equal variant and Mann-Whitney tests when the data was non-normal or did not have homogeneity of variance. Equivalency of the design problems to the peanut problem was investigated using the Mann-Whitney test for equivalency [81, 82] as outlined in section 2.6. Additionally, the confidence interval inclusion procedure [82] was run to visualize the mean differences.

Table 3. Between-Subjects Problem Sample Sizes

Section	Peanut	Corn	Coconut	Alarm	Total:
A	11	8	8	8	35
B	11	12	12	12	47
D	9	10	8	7	34
G	9	8	9	9	35
I	10	10	9	10	39
Total:	50	48	46	46	190

4.1.1 Section Comparison

Initial analysis was done to check for significant differences in the data resulting from outside variables. For example, an outside variable may include the effect of testing

section time on the solutions generated. As the similarities between problems were not yet investigated, the collection sessions were compared for each design problem. The mean data was investigated for normality for each of the individual groups and was investigated for equal variance between the test groups using a Shapiro-Wilks and Levene's Test respectively. The data was then tested for differences using an ANOVA. Since it was found that not all pairings had normality and equal variance, a Kruskal-Wallis H test was used in the comparison of the groups. The results can be seen in Table 4.

Table 4. Section Comparison by Problem

Section Comparison		ANOVA		Kruskal-Wallis		Levene's Test		All Sections Normal
		F	Sig.	Chi-Square	Asymp. Sig.	Levene Statistic	Sig.	
Peanut	**Quantity**	0.70	0.60	2.91	0.57	0.65	0.63	Yes
	Quality	0.56	0.69	2.36	0.67	0.33	0.85	Yes
	Novelty	0.79	0.54	2.95	0.57	0.29	0.88	Yes
	Variety	3.03	0.03	9.01	0.06	1.56	0.20	**No**
	Number of Solutions	2.47	0.06	7.36	0.12	2.02	0.11	Yes
Corn	**Quantity**	0.17	0.95	0.81	0.94	0.54	0.71	Yes
	Quality	1.13	0.36	5.43	0.25	1.27	0.30	**No**
	Novelty	0.55	0.70	4.52	0.34	1.50	0.22	**No**
	Variety	0.14	0.97	0.86	0.93	1.93	0.12	Yes
	Number of Solutions	0.72	0.59	2.86	0.58	1.31	0.28	**No**
Coconut	**Quantity**	0.62	0.65	2.09	0.72	2.60	0.05	Yes
	Quality	1.83	0.14	6.77	0.15	0.89	0.48	Yes
	Novelty	2.82	0.04	9.12	0.06	0.81	0.52	Yes
	Variety	1.62	0.19	4.42	0.35	5.15	0.00	Yes
	Number of Solutions	1.10	0.37	2.05	0.73	2.46	0.06	Yes
Alarm	**Quantity**	0.31	0.87	1.08	0.90	0.92	0.46	Yes
	Quality	0.66	0.63	3.55	0.47	2.62	0.05	Yes
	Novelty	0.37	0.83	1.21	0.88	0.61	0.66	Yes
	Variety	0.36	0.83	2.25	0.69	0.38	0.83	**No**
	Number of Solutions	0.32	0.86	1.94	0.75	0.73	0.58	**No**
Key	**Statistically Different**		**Marginal Difference**		**Equal Variant**			

As can be seen in the table, for the peanut problem there is a significant difference ($p < .05$) between the sessions with respect to the variety metric and the for the coconut problem there is a significant difference in the novelty metric according to the ANOVA test. However, upon closer inspection, the variety metric does not meet the assumption of equal variance, therefore the Kruskal-Wallis H test should be looked at instead. Although close, it can be seen that it does not show a statistically significant difference ($p < .05$) between the sections. When looking at the novelty metric for the coconut problem the assumptions of the ANOVA were met, however, the significance ($F=2.820$, $p=0.037$) coupled with the significance of the Kruskal-Wallis H test ($X^2=9.117$, $p=0.058$) suggests that the significance may be borderline. The data was graphed for both instances where significance was detected and can be seen in Figure 1 where the error bars present represent standard error. From the figure, it can be seen that the variations across for the novelty metric of the coconut problem appear to be random fluctuations and not a particular pattern. The significance of the data was therefore considered sufficiently similar by the author to combine the sessions into larger groups for more powerful analysis.

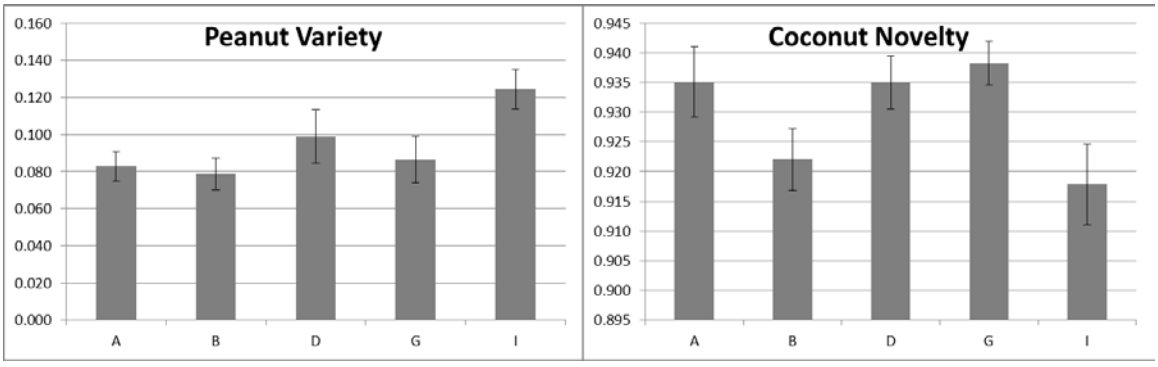

Figure 1. Significant Section Comparisons

A final outside variable needed to be checked before the data could be combined to the problem level [97]. As part of a different study involving the same participant group, sections A, B, D, and I were taught by a single professor and two section groups (Sections A and B) were taught an additional ideation technique, Design Heuristics, over the course of a semester. The remaining section (Section G) was taught by another professor where the material taught was the same as that in sections D and I.

Design Heuristics is a method of idea generation where users are given a set of cards where each card has a unique design concept that can be used as inspiration in the design process [98]. For this method of ideation to be most effective, the users must have the cards present during ideation, which was not the case in this study because all participants were prevented from using outside materials. The introduction of Design Heuristics over the course of the semester could still bias the data, however, as participants could remember particular cards during the ideation session. The affect was therefore investigated between the four sections taught by the same professor with the traditional method (Sections D and I) considered the control and the remaining sections (Sections A and B) considered the treatment.

Similarly to the section-based analysis, the groups were first checked for normality and equal variance using the Shapiro-Wilks and Levene's Test respectively. Since there were only two groups in comparison now (sections A and B, and sections D and I), an independent samples t-test was used for analysis as well as a Mann-Whitney U Test to account for instances when the assumptions for the t-test were not met. The comparison was again made within each problem, as the relationship between the groups was not yet investigated. The results of this comparison can be seen in Table 5.

Table 5. Design Heuristics Comparison

Treatment Comparison		Independent Samples t-test		Mann-Whitney U		Levene's Test		All Groups Normal
		t	Sig. (2-tailed)	U	Asymp. Sig. (2-tailed)	Levene Statistic	Sig.	
Peanut	Quantity	1.52	0.14	149.50	0.12	0.14	0.71	Yes
	Quality	1.36	0.18	157.50	0.17	0.00	0.98	Yes
	Novelty	1.34	0.19	163.00	0.23	0.39	0.53	Yes
	Variety	3.01	0.00	120.00	0.02	5.61	0.02	**No**
	Number of Solutions	2.42	0.02	117.00	0.01	0.14	0.71	**No**
Corn	Quantity	-0.17	0.87	196.50	0.92	0.21	0.65	Yes
	Quality	0.54	0.59	170.00	0.41	0.03	0.86	**No**
	Novelty	-1.32	0.20	143.50	0.13	0.08	0.78	Yes
	Variety	-0.79	0.44	176.50	0.52	1.69	0.20	Yes
	Number of Solutions	-1.43	0.16	160.50	0.28	2.66	0.11	**No**
Coconut	Quantity	1.20	0.24	142.00	0.39	3.02	0.09	Yes
	Quality	0.65	0.52	147.00	0.48	0.32	0.58	Yes
	Novelty	-0.22	0.83	165.50	0.89	0.03	0.86	Yes
	Variety	1.15	0.26	159.00	0.73	8.29	0.01	**No**
	Number of Solutions	0.87	0.39	153.50	0.61	2.24	0.14	**No**
Alarm	Quantity	-0.98	0.33	147.00	0.48	0.82	0.37	Yes
	Quality	1.38	0.18	139.00	0.34	2.99	0.09	Yes
	Novelty	-0.36	0.72	161.50	0.80	1.62	0.21	Yes
	Variety	-0.39	0.70	169.00	0.98	1.05	0.31	**No**
	Number of Solutions	-0.31	0.75	166.50	0.91	0.87	0.36	Yes
Key	**Statistically Different**		**Marginal Difference**		**Equal Variant**			

As can be seen in the table, there is statistically significant differences in the variety and number of solutions metrics for the peanut problem. This says that the group not exposed to the treatment (design heuristics) produced more solutions and solutions of a higher variety than those exposed to the treatment. To get a more complete understanding of the data, the data was graphed and can be seen in Figure 2.

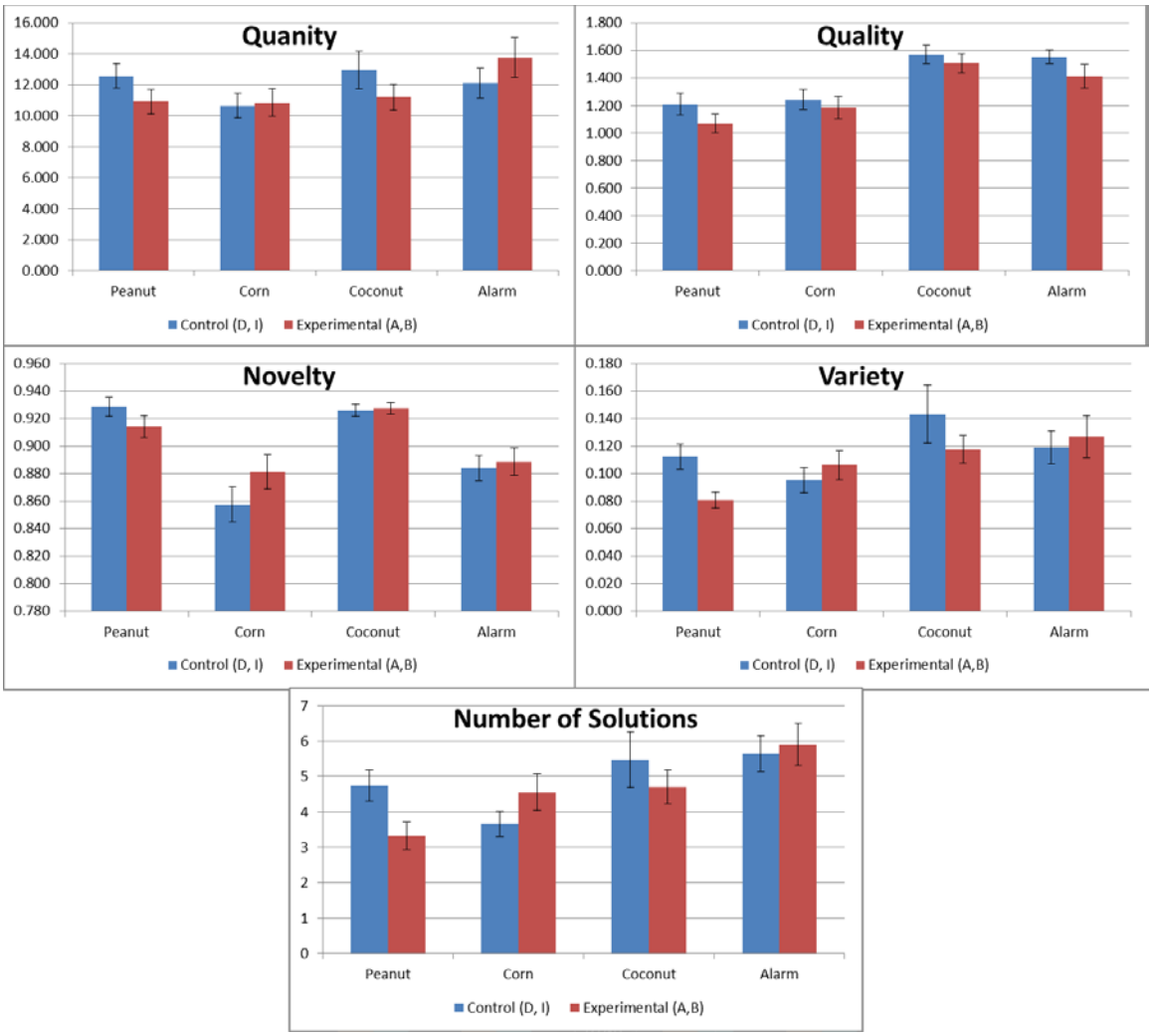

Figure 2. Design Heuristics Comparison

Upon closer inspection of the data, the statistical differences is most likely caused from the control group in one class section, Section I, performing higher in these metrics. This performance is what was driving the differences seen in the peanut problem when looked at previously. As this difference was only seen in the peanut problem and the cause most likely traced back to an individual class and not the treatment, the data was considered by the author to be unaffected by the treatment. The data was therefore combined across all sections for each individual problem.

4.1.2 Problem Differences Comparison

At this stage of the analysis, the data for each problem was again run through normality and homogeneity of variance testing using Shapiro-Wilks and Levene's Test respectively. The differences between the problems were then tested across all problems using an ANOVA and Kruskal-Wallis test to account for cases where the assumptions for an ANOVA were violated. The resulting analysis can be seen in Table 6.

Table 6. Between-Subject Comparison Across the Different Problems

Problem Comparison	ANOVA		Kruskal Wallis				All Groups Normal
	F	Sig.	Chi-Square	Asymp. Sig.	Levene Statistic	Sig.	
Quantity	2.25	0.08	5.94	0.11	0.62	0.60	Yes
Quality	19.94	0.00	51.73	0.00	1.01	0.39	**No**
Novelty	21.51	0.00	48.25	0.00	14.80	0.00	Yes
Variety	5.77	0.00	14.83	0.00	3.43	0.02	**No**
Number of Solutions	6.56	0.00	21.26	0.00	0.81	0.49	**No**
Key	**Near to Statistical Difference for Mean Number**				**Equal Variant**		

In this instance, it was of particular interest if the data was not statistically different. It was found that across all four problems tested, the least amount of difference (F=2.250, p=0.084) was found in the quantity metric. This indicates that there are not similarities across all the problems, which is not surprising. When looking at the other metrics, it was found that the problems behaved statistically differently. To understand the source of the differences, the data was graphed comparing the design problems for each metric and can be seen in Figure 3.

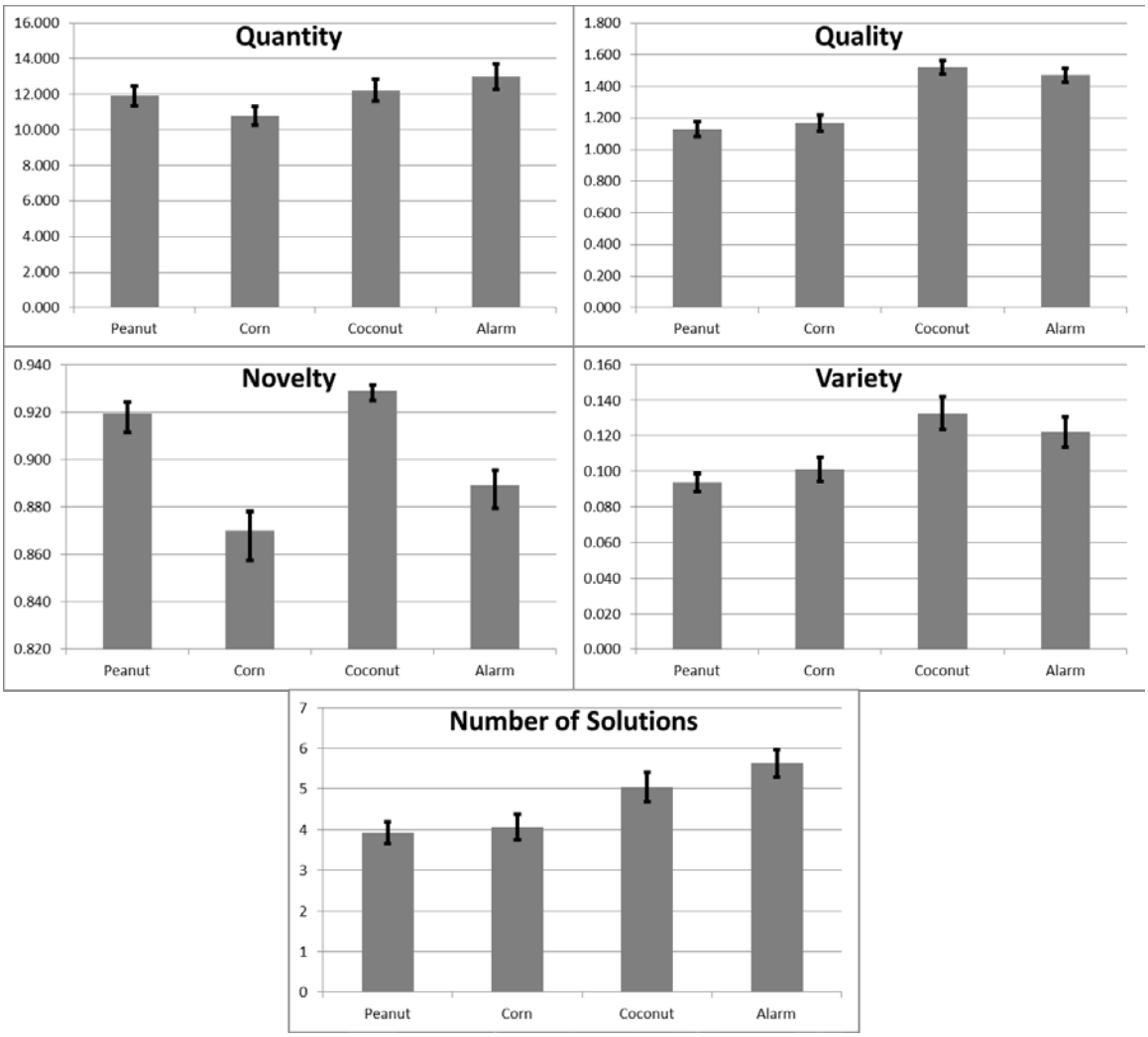

Figure 3. Problem Comparisons

The figure demonstrates that there are differences between some of the problems tested, but it also shows that some groupings of problems have more similarities than expressed by the original statistics. As the research is aiming to understand the relationships between the problems, the problems were compared individually using an independent samples t-test and Mann-Whitney U test to account for cases where the assumptions of a t-test were not met. As the peanut problem was selected as the

benchmark design problem, comparisons made to the peanut problem are of particular interest to the research and are documented in Table 7.

Table 7. Peanut Problem Comparisons

Peanut Comparisons		Independent Samples t-test		Mann-Whitney U		Levene's Test		All Groups Normal
		t	Sig. (2-tailed)	U	Asymp. Sig. (2-tailed)	Levene Statistic	Sig.	
Peanut-Corn	Quantity	1.45	0.15	1001.50	0.16	0.05	0.82	Yes
	Quality	-0.58	0.56	1154.50	0.74	0.03	0.87	**No**
	Novelty	5.20	0.00	569.00	0.00	8.70	0.00	Yes
	Variety	-0.86	0.39	1086.00	0.42	2.88	0.09	**No**
	Number of Solutions	-0.35	0.72	1198.00	0.99	0.75	0.39	**No**
Peanut-Coconut	Quantity	-0.39	0.70	1095.00	0.69	0.55	0.46	Yes
	Quality	-6.41	0.00	399.00	0.00	1.64	0.20	Yes
	Novelty	-1.69	0.10	980.00	0.21	21.38	0.00	Yes
	Variety	-3.74	0.00	735.00	0.00	7.11	0.01	**No**
	Number of Solutions	-2.53	0.01	821.00	0.01	1.75	0.19	**No**
Peanut-Alarm	Quantity	-1.22	0.23	994.50	0.25	1.31	0.26	Yes
	Quality	-5.39	0.00	452.50	0.00	1.33	0.25	**No**
	Novelty	3.89	0.00	677.00	0.00	2.21	0.14	Yes
	Variety	-2.88	0.00	855.00	0.03	9.58	0.00	**No**
	Number of Solutions	-4.04	0.00	616.00	0.00	2.04	0.16	**No**
Key	**Not Statistically Different**			**Marginal Difference**		**Equal Variant**		

The table demonstrates that there is a large amount of similarity between the peanut and corn problems when looking at group means. According to the data, there is no statistical difference between these two problems for all metrics except for novelty. However, this does not mean that these problems are equivalent, but rather that there is not enough evidence to say they are different. The lack of differences between the peanut and corn problems suggests that they are candidates to be used to test differences based on group means but further equivalency tests would be required to confirm this. Additionally, the problems may exhibit differences in within-subjects testing as the similarities may be only at a group level and individuals do not respond the same way.

Table 7 also shows the relationship of the peanut problem to the coconut and alarm problems. The peanut problem continued to demonstrate similarity to the coconut and alarm problems when graded on the quantity metric and similarity with the coconut problem when graded on the novelty metric. This similarity was close to the threshold however, so should be viewed accordingly. The remaining problem comparisons (corn-alarm, corn-coconut, and alarm-coconut) were analyzed to determine other problem similarities, which can be seen in Table 8.

Table 8. Other Problem Comparisons

Other Comparisons		Independent Samples t-test		Mann-Whitney U		Levene's Test		All Groups Normal
		t	Sig. (2-tailed)	U	Asymp. Sig. (2-tailed)	Levene Statistic	Sig.	
Corn-Coconut	Quantity	-1.75	0.08	880.50	0.09	0.31	0.58	Yes
	Quality	-5.44	0.00	452.00	0.00	1.68	0.20	No
	Novelty	-6.69	0.00	309.00	0.00	34.32	0.00	Yes
	Variety	-2.75	0.01	738.00	0.01	1.56	0.21	No
	Number of Solutions	-2.06	0.04	813.00	0.03	0.29	0.59	No
Corn-Alarm	Quantity	-2.45	0.02	807.50	0.02	0.98	0.32	Yes
	Quality	-4.51	0.00	511.00	0.00	1.41	0.24	No
	Novelty	-1.86	0.07	903.00	0.13	2.97	0.09	Yes
	Variety	-1.92	0.06	790.00	0.02	1.98	0.16	No
	Number of Solutions	-3.45	0.00	628.00	0.00	0.30	0.58	No
Alarm-Coconut	Quantity	0.81	0.42	971.50	0.50	0.24	0.63	Yes
	Quality	-0.85	0.40	953.00	0.41	0.00	0.99	No
	Novelty	-5.99	0.00	425.00	0.00	33.93	0.00	Yes
	Variety	-0.85	0.40	860.00	0.12	0.00	0.98	No
	Number of Solutions	1.18	0.24	860.50	0.12	0.00	0.97	No
Key	Not Statistically Different			Marginal Difference		Equal Variant		

Some interesting trends emerged in Table 8. It was seen that the similarity in quantity metric scores was consistent for the corn-coconut and alarm-coconut parings but not for the corn-alarm pairing. The corn and alarm problems did show threshold similarity on the novelty metric. Perhaps the most interesting find in the remaining

comparisons is the similarities present between the alarm and coconut problems. These two problems were not different in all metrics except for the novelty metric, which are the same metrics that showed no difference with the peanut and corn problems.

By looking at the differences between the design problems, we are able to gain a more complete picture of the relationships that they share. For example, it was shown that the quantity metric was not statistically different between most of the problem pairings. Although this does not mean equivalence for the quantity metric, it may indicate an outside variable affecting the quantity metric. For example, there may be a relative limit of features generated over this time span, which all participants are able to reach. It may also suggest that the intervention methods taken, namely rewording the problems and providing a consistent number of customer needs, leads the participants to produce similar numbers of features over the ideation session. The similarities of the peanut and corn problems, and the alarm and coconut problems suggest that there are certain outside variables that may affect the ideation results. A likely source of these similarities is the presence of similar problem characteristics. These characteristics were hypothesized in previous literature [3] and believed to influence research outcomes. As these characteristics were previously believed to have an impact, three of these characteristics suspected of having the largest impact are investigated in this study and results are shown in section 4.1.5.

4.1.3 Equivalency Results

As discussed in section 2.6, while traditional statistical testing such as the t-test and Mann-Whitney U test used in the previous section are beneficial in understanding

what differences exist between groups, they cannot by the nature of their design assert equivalency between groups. For this reason, the problems were tested for equivalency using the Mann-Whitney test for equivalence with equivalence limits of $\varepsilon = \pm 0.2$ and a significance of $\alpha = 0.05$. As the peanut problem was selected as the baseline for the study, the remaining design problems were compared to it for each metric. The results from this can be seen in Table 9 below. As discussed in section 2.6.2, equivalency is assessed using this methodology by evaluating if a variable, T, is less than a critical value C_{MW}. The variable T represents the left side of Equation (5) and is dependent on the U-statistic estimator W_+ of the data, as well as the equivalence limits, ε. The variable C_{MW} is the right side of Equation (5), and is dependent on the equivalence limits, ε, and the significance level, α, as seen in Equation (6).

Table 9. Equivalence Testing Results

		W+	$\sigma[W_+]$	T	C_{MW}	Equivalent
Peanut-Corn	**Quantity**	0.545	0.057	0.792	0.101	No
	Quality	0.413	0.058	1.515	0.101	No
	Novelty	0.763	0.047	5.602	0.128	No
	Variety	0.453	0.058	0.820	0.100	No
	Number of Solutions	0.418	0.057	1.433	0.102	No
Peanut-Alarm	**Quantity**	0.399	0.057	1.757	0.101	No
	Quality	0.173	0.042	7.765	0.152	No
	Novelty	0.706	0.051	3.997	0.114	No
	Variety	0.372	0.056	2.284	0.103	No
	Number of Solutions	0.205	0.045	6.539	0.136	No
Peanut-Coconut	**Quantity**	0.440	0.058	1.045	0.100	No
	Quality	0.145	0.037	9.501	0.193	No
	Novelty	0.426	0.059	1.248	0.098	No
	Variety	0.320	0.054	3.346	0.108	No
	Number of Solutions	0.284	0.051	4.205	0.114	No

As can be seen, the Mann-Whitney test for equivalence revealed that there was no statistical equivalence found between any of the design problems at this significance level for these equivalence limits. However, it continues to demonstrate the potential of the peanut and corn problems as candidates for problem equivalency as they are again the closest problems tested to equivalency. As this test looks at both the raw difference as well as the distribution differences in the data, the test was rerun with a linear shift applied to each metric to understand what was causing the inequality. In this way, if the test now resulted in equivalence, the primary difference was the raw scores, which can be compensated for in future analysis between those metrics. However, if the test again reported that there is insufficient evidence for equivalency in the metrics, than the inequality can be attributed to different distributions in the scores across the design problems. The results of this test can be seen below in Table 10.

Table 10. Equivalency Testing Results after Linear Shift

		W_+	$\sigma[W_+]$	T	C_{MW}	Equivalent
Peanut-Corn	**Quantity**	0.470	0.058	0.528	0.101	No
	Quality	0.550	0.058	0.857	0.100	No
	Novelty	0.470	0.059	0.501	0.099	No
	Variety	0.512	0.061	0.193	0.096	No
	Number of Solutions	0.580	0.058	1.385	0.101	No
Peanut-Alarm	**Quantity**	0.535	0.058	0.596	0.100	No
	Quality	0.489	0.059	0.192	0.099	No
	Novelty	0.493	0.059	0.126	0.099	No
	Variety	0.535	0.060	0.588	0.097	No
	Number of Solutions	0.481	0.059	0.317	0.099	No
Peanut-Coconut	**Quantity**	0.513	0.058	0.216	0.099	No
	Quality	0.479	0.059	0.364	0.099	No
	Novelty	0.515	0.061	0.244	0.096	No
	Variety	0.534	0.060	0.564	0.097	No
	Number of Solutions	0.588	0.058	1.506	0.100	No

As can be seen in Table 10 the equivalency test still maintains that the problems are not statistically equivalent while having the same mean for any of the metrics. However, the data is much closer to equivalency for a number of the metrics tested. This means that the distribution differences in the data are causing the inequalities which leads to two observations. First, it shows that linear equating is not a sufficient method on its own of making two problems equivalent. Secondly, as the inequality is caused by a distribution difference, it may be possible that the problems could be equivalent with larger sample sizes because at sufficient sample sizes, the distribution may tend to a Gaussian distribution. Further research could observe the effect with larger sample sizes.

4.1.4 Confidence Interval Inclusion

In order to investigate the relationships between the raw mean differences, the author employed the confidence interval inclusion method. This method serves as a means to visualize the differences in this study, and not to assess equivalency due to the nonparametric nature of the data. Equivalency assessments in this study for between-subjects data use the Mann-Whitney test for equivalency, seen in the previous section. Additionally, the confidence interval inclusion method is presented to demonstrate its capabilities and interpretations in order to build upon a framework for testing for equivalent design problems.

The equivalence interval used in this study was established according to the same guidelines used previously, corresponding to a ± 20% interval, by using an equivalence limit ε' of 0.5 as specified by Wellek [82]. The interval is symmetric about the mean with respect to the standard deviation of the peanut problem. The resulting equivalence

intervals and confidence intervals for the problem pairings are available in Table 11 and graphically in Figure 4.

Table 11. Confidence Interval Inclusion Data

	Equivalency Interval		Peanut-Corn		Peanut-Alarm		Peanut-Coconut	
			90% Confidence Interval		90% Confidence Interval		90% Confidence Interval	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
Quantity	-1.90	1.90	-0.16	2.38	-2.55	0.40	-1.67	1.03
Quality	-0.16	0.16	-0.15	0.07	-0.45	-0.24	-0.49	-0.29
Novelty	-0.02	0.02	0.03	0.07	0.02	0.04	-0.02	0.00
Variety	-0.02	0.02	-0.02	0.01	-0.04	-0.01	-0.06	-0.02
Number of Solutions	-0.93	0.93	-0.81	0.53	-2.41	-1.01	-1.86	-0.38

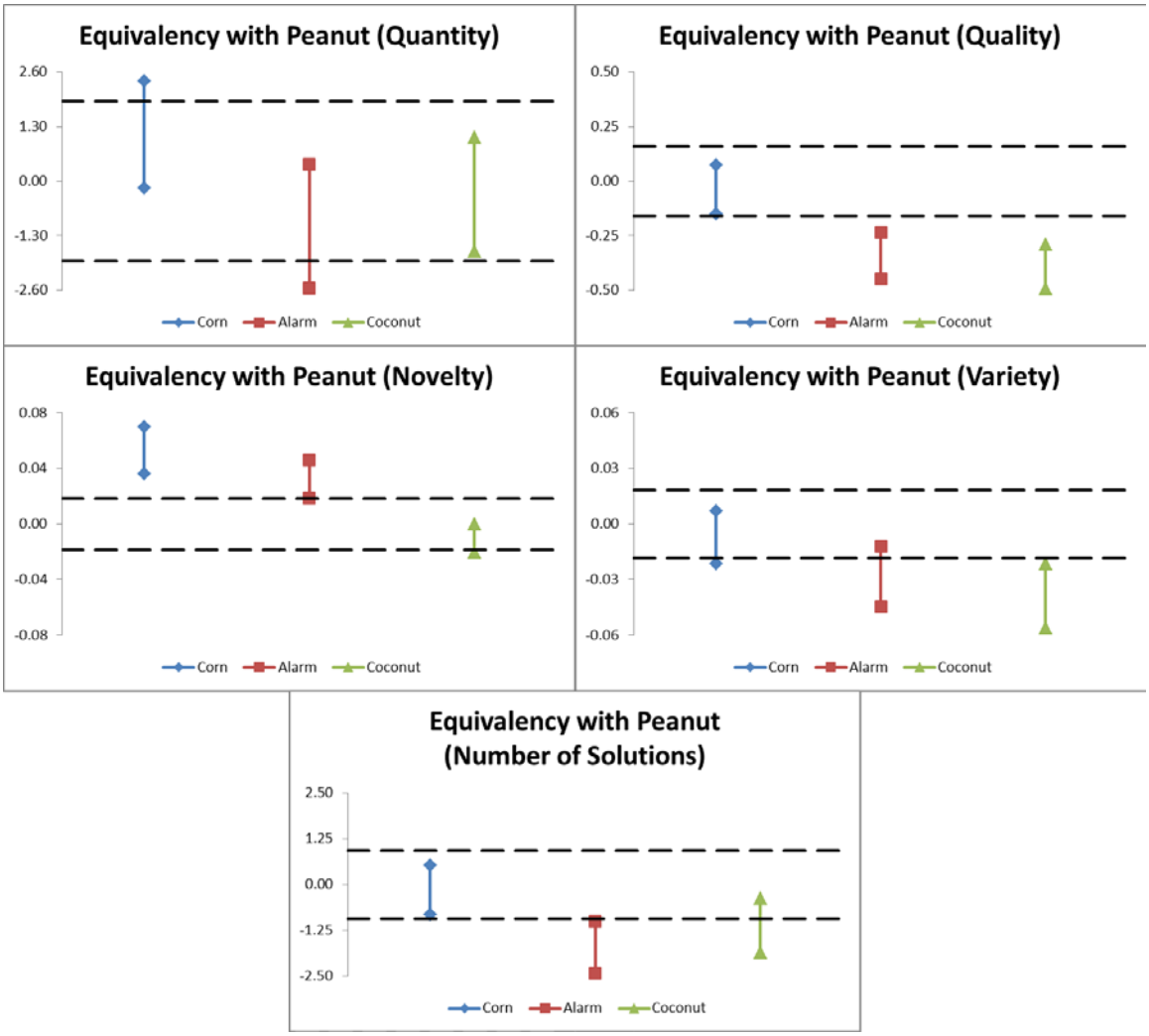

Figure 4. Confidence Interval Inclusion

At first glance, the figure appears to show equivalency for some metrics, such as peanut-corn (number of solutions) as the confidence interval is included within the equivalence limits. However, this equivalency can be misleading due to the nonparametric nature of the data. The graph does demonstrate other trends that may exist in the data, however. For example, a number of confidence intervals, such as those associated with the variety metric, are approximately the same size as their corresponding equivalence intervals, which suggests that a larger sample size would be necessary to

draw conclusions. Additionally, significant differences occur in the figure anywhere that the confidence interval does not cross zero, which gives a visual comparison to the difference results from section 4.1.2. Confidence interval inclusion can also be used in order to determine if certain problem pairings would benefit from linear mapping. In this case, the confidence intervals for the quantity metric are small enough that a linear shift applied to the data may be used to attain equivalency if the distributions are the same between the problems.

4.1.5 Survey Correlations

In addition to the metrics specified earlier in the paper, the students in the between-subjects group were given a survey after the ideation session to gather additional information on problem characteristics. The target characteristics were gathered from literature [3] and include: problem difficulty, familiarity with the design problem, and familiarity with the design solution. All of these characteristics were assessed with self-reported data using a 5-point scale with the question format available in Appendix C. The data collected was graphed for each problem and can be seen in Figure 5.

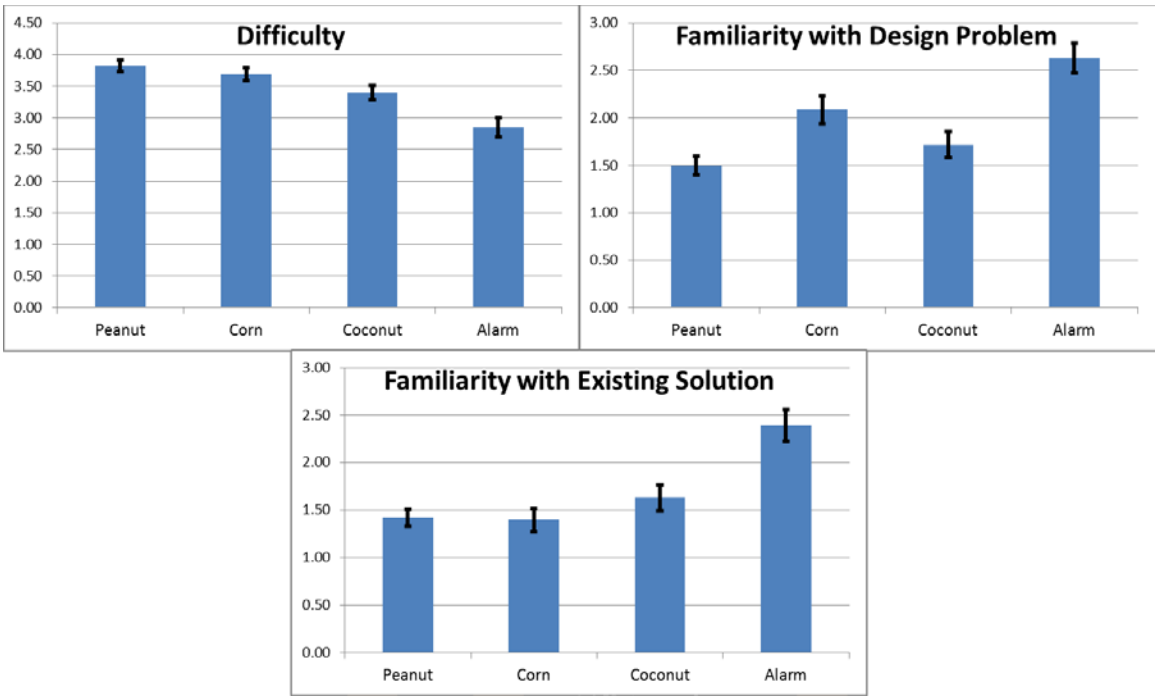

Figure 5. Problem Characteristics

As can be seen from the figure, the difficulty of the problem is inversely proportional to the familiarity with existing solutions. The same trend is observed between difficulty and familiarity with the design problem, just not as pronounced. The data was processed in the same manner as the between-subjects data and compared across all problems and then individually for each problem pairing. The data was found to be non-normal; therefore, non-parametric tests were utilized. There were significant differences for these characteristics between all problems, but when investigated in pairs some problems demonstrated similar characteristics, as shown in Table 12.

Table 12. Problem Characteristics Comparison

Problem Characteristics		Mann-Whitney U	
		U	Asymp. Sig. (2-tailed)
Peanut-Corn	Difficulty of Problem	1095.500	0.404
	Familiarity with Problem	799.000	0.002
	Familiarity with Existing Solutions	1100.000	0.370
Peanut-Coconut	Difficulty of Problem	828.500	0.009
	Familiarity with Problem	1035.500	0.344
	Familiarity with Existing Solutions	1059.500	0.436
Peanut-Alarm	Difficulty of Problem	542.500	0.000
	Familiarity with Problem	456.500	0.000
	Familiarity with Existing Solutions	571.000	0.000
Corn-Coconut	Difficulty of Problem	898.500	0.088
	Familiarity with Problem	863.000	0.052
	Familiarity with Existing Solutions	935.500	0.118
Corn-Alarm	Difficulty of Problem	603.500	0.000
	Familiarity with Problem	783.000	0.012
	Familiarity with Existing Solutions	525.000	0.000
Alarm-Coconut	Difficulty of Problem	728.500	0.007
	Familiarity with Problem	551.000	0.000
	Familiarity with Existing Solutions	635.500	0.000
	Key	Not Statistically Different	

Due to the similarities seen previously between the peanut and corn problems and the coconut and alarm problems, these pairings were of particular interest. The peanut and corn problems showed similarity in problem difficulty (U=1095.5, p=0.404) and familiarity with existing solutions (U=1100.0, p=0.370), however, when the alarm and coconut problems were compared all three problem characteristics were found to be statistically different. This indicates that although these characteristics may be important, their correlations between problems may not necessarily lead to problem differences. Outside of problem comparisons, the correlation of these characteristics with the different metrics was investigated to see if they had an impact, with the results shown in Table 13.

Table 13. Correlation of Problem Characteristics and Ideation Metrics

All	Spearman's rho		
	Difficulty	Problem Familiarity	Solution Familiarity
Quantity	**-.166***	**0.107**	**.179***
Quality	**-.195****	**0.129**	**.157***
Novelty	**0.088**	**-.157***	**-0.012**
Variety	**-0.122**	**0.010**	**0.111**
Solutions	**-.220****	**0.043**	**.177***
Key	**Slight (0.1-0.29)**	**Moderate (0.3-0.49)**	**Strong (0.5-1)**
**. Correlation is significant at the 0.01 level (2-tailed).		*. Correlation is significant at the 0.05 level (2-tailed).	

The table shows that there are a number of correlations present between the design characteristics and the ideation metrics. That being said, the correlations seen are only slight correlations throughout, but may speak to some trends between the characteristics and metrics. It is seen that the more difficult the problem, the lower the metrics scores and the more familiar the participant is with existing solutions, the higher the metrics scores. The results demonstrate that there may indeed be correlations between the metrics and the problem characteristics but the metrics selected for this study did not have the largest impact. It may also mean that the correlations are more complex and have to do with the interactions between multiple characteristics and not dominated by a single characteristic.

4.2 Within-Subjects Testing

The same metric grading procedures were used for the within-subjects data as the between-subjects data. Once graded, the data was first compared on a group mean basis to see if the same differences seen in the between-subjects data between the peanut and corn problems were present. Normality and homogeneity of variance between the problems were checked with a Shapiro-Wilks and Levene's Test respectively. The data

was then compared using a paired samples t-test and Wilcoxon signed ranks test for cases where the assumptions for the paired t-test were not met and can be seen in Table 14.

Table 14. Within-Subjects Mean Comparison

Peanut-Corn	Wilcoxon Signed Ranks		Paired Samples t-test		
	Z	Asymp. Sig. (2-tailed)	t	df	Sig. (2-tailed)
Quantity	-1.616[b]	0.106	1.683	38	0.101
Quality	-.359[b]	0.719	0.535	38	0.596
Novelty	-3.586[b]	0.000	4.489	38	0.000
Variety	-.503[c]	0.615	-0.974	38	0.336
Number of Solutions	-.787[b]	0.431	0.461	38	0.647
Key	b. Based on positive ranks.		c. Based on negative ranks.		
	Not Statistically Different				

The table demonstrates that the same trends seen previously between the peanut and corn data remains true using this different data set. This is encouraging, as it builds upon the between-subjects data in what metrics are significantly different between these problems. The correlation of the students' scores on each metric across the problems was compared using Spearman's correlation to account for the data being non-normal with the results shown in Table 15. The correlations were evaluated according to guidelines put forth by Cohen [99] with a light correlation for values ranging from 0.1-0.3, a moderate correlation for values ranging from 0.3-0.5, and a strong correlation for values greater than 0.5.

Table 15. Within Subjects Spearman's Correlation

Peanut-Corn		Quantity	Quality	Novelty	Variety	Number of Solutions
Correlation	Spearman's Rho	.508	0.227	0.011	.328	.683
	Sig. (2-tailed)	0.001	0.164	0.945	0.041	0.000
Key	Slight(0.1 - 0.29)		Moderate (0.3 - 0.49)		Strong (0.5 - 1)	

The table shows some interesting results in the data. It can be seen that there is some level of positive correlation in the quantity, quality, variety, and number of solutions metrics, which is every metric that was shown as similar during group mean testing. This indicates that the between-subjects testing may be a quick indicator of what metrics will correlate at the individual level. That being said, only the quantity and number of solutions metrics show strong correlations (ρ=0.508 and ρ=0.683 respectively), while the variety metric shows moderate correlation (ρ=0.328), and the quality metric has a slight correlation (ρ=0.227).

The strong correlations present in the quantity and number of solutions metrics are extremely valuable to the field. This indicates that these problems can be used in the form presented in this research for with-subject analysis for quantity and number of solutions. This therefore represents the first set of problems empirically shown to be equivalent on ideation metrics. Of particular importance to the field is the fact that the quantity metric is shown as one of the equivalent metrics as this is the most widely used metric in design [1, 14, 18] and in looking at creativity in general [2]. The data was also expressed graphically to check for trends that may be hidden in the data. This can be seen for each metric in Figure 6 with the participants graphed on the principal axis.

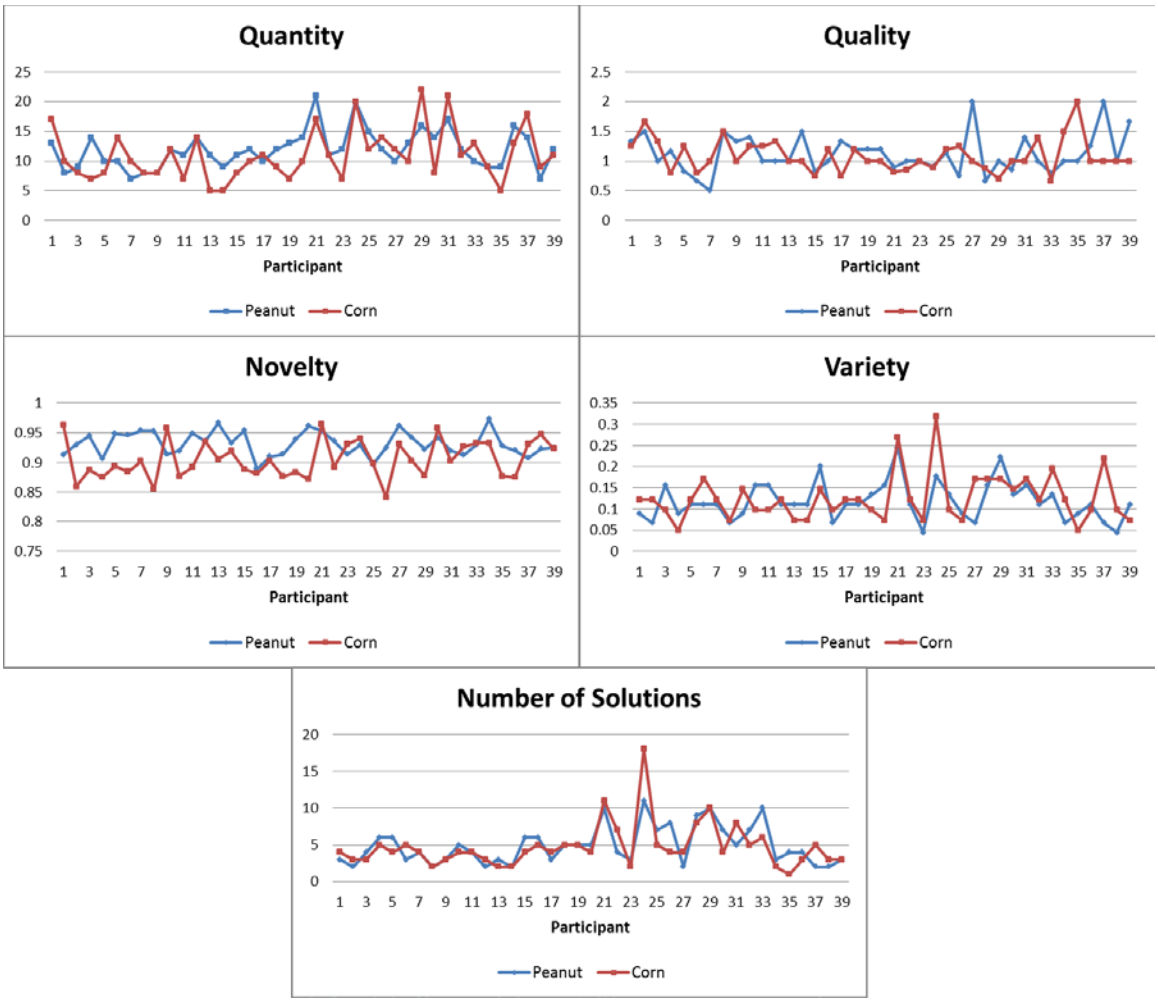

Figure 6.Within-Subjects Correlation

The data was additionally sorted on various qualifiers, such as quantity scores for the peanut problem and the data visualized. An example of checking for trends can be seen in where the data was sorted according to the scores of the participants on the peanut design problem. This was repeated for all metrics according to each design problem as well as the difference in scores between the design problems. Through this investigation, no noticeable trends were observed.

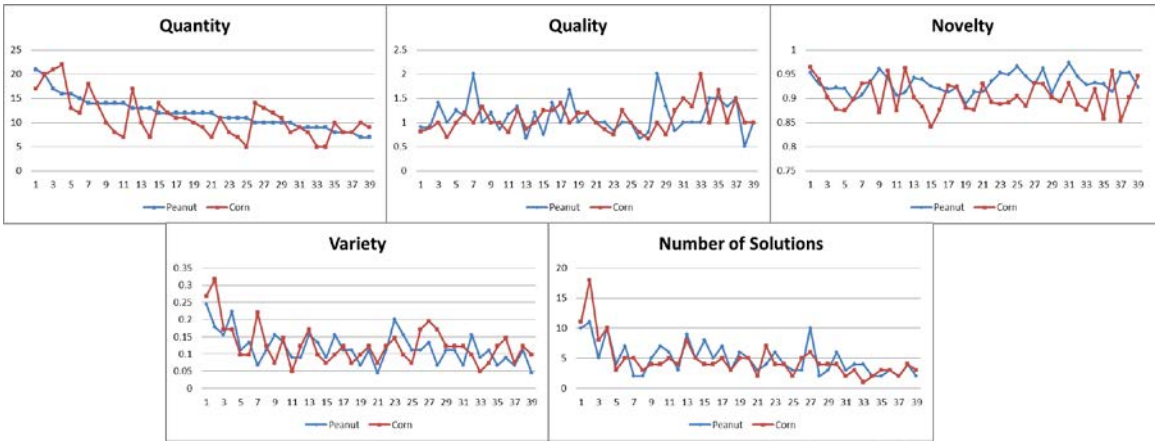

Figure 7. Within-Subjects Correlation sorted on Peanut Quantity

4.3 Problem Equivalency Framework

As a result of this research, a framework has been developed for use in testing for problem equivalency. The framework, presented in this section in the form of guidelines, should be used in future testing in order to evaluate design problems for equivalency. This methodology can also be used to test changes to design problems for increased equivalency. The framework presented allows a pre-screening process followed by a more in-depth equivalency testing at the within-subject level. While not necessary to demonstrate problem equivalency, it is the recommendation of the author that a characteristics survey like the one shown in this study is conducted as part of the within-subjects testing. This would provide the field with valuable insight into how differences in problem characteristics directly influence problem equivalency.

4.3.1 Equivalency Pre-Screening

The process of testing for equivalency can be somewhat demanding due to the within-subject testing required. In order to help alleviate this problem between-subjects

approaches are presented as a means to pre-screen design problems for equivalency. The pre-screening process works by first dividing a sample population into a number or groups equal to the number of problems being screened. Each group is then asked to individually generate solutions for their design problem. The solutions are scored across ideation metrics and the results compared. Significant differences, or lack thereof, can be used as an initial measure of problem pairings that are good candidates for further study. Additionally, the confidence interval inclusion method can be used on parametric data to test for initial equivalency. The confidence interval inclusion method also visually shows which problems may be good candidates for mapping scores from one problem to the next if correlations are present. This would be the case when the confidence interval of the mean difference is much smaller than the equivalence limits but does not contain zero. Nonparametric data can be tested using the Mann-Whitney test for equivalence to determine initial equivalencies. As this test looks at distribution differences as well as mean differences it can be used to pinpoint the source of any in equivalencies found. Using this method, problem pairings that show initial similarities based on lack of statistical differences are selected for further study, with the source of any inequalities traced to distribution differences or mean differences based on the confidence interval inclusion and/or Mann-Whitney test for equivalence.

4.3.2 Equivalency Testing

Once selected from pre-screening, design problems are tested for equivalency using within-subjects design. Within-subject testing allows analysis according to correlation values, which is the methodology required for equivalency of parallel forms [79]. While the test setup for this within-subject analysis can be carried out using two or

more problems, the methodology used for two problems will be demonstrated here for simplicity. Researchers should first divide the participants into a number of groups equal to the factorial of the number of design problems under investigation. In this example there would be two groups, while if there were three problems than there would be six groups. Participants in each group would be given a different design problem and asked to individually generate solutions. For this study, an idea generation session lasting 50 minutes was used without the assistance of outside stimuli during the design process. The process is then repeated, assigning different design problems to each group, according to how many problems are being tested. In this example, two problems would use two idea generation session, while three problems would use three idea generation sessions. Between each idea generation session sufficient time should be given to reduce fatigue, in this study a period of two days between sessions was used.

Once all sessions are complete, researchers can evaluate the solutions according to ideation metrics. The scores should be analysed first for differences in the data brought about by run order using traditional statistical methods such as an ANOVA. The data can then be analysed using correlations, Pearson's for parametric and Spearman's for nonparametric. Correlation values of 0.5 and higher represent sufficient correlation values to state equivalency in the design problems, according to the metrics under investigation.

CHAPTER 5. CONCLUSIONS

Over the course of the study, four design problems were studied in order to determine if the problems could be considered as equivalent for the purpose of design studies. Through within-subject analysis, it was shown that the peanut and corn problems were equivalent for two metrics, quantity and number of solutions. This means that at this stage researchers in design can use these problems for within-subject experiment. This represents a significant accomplishment for researchers, as to this point there has not been empirically determined highly similar problems. The presence of equivalent problems will allow researchers to expand their testing procedures to include within-subjects analysis, paving the way for high validity investigations.

This work also presents a framework for seeking equivalent design problems. In addition to within-subjects analysis, a between-subjects analysis was used to determine whether between-subject testing was a viable option for determining equivalency. During this part of the research, it was found that the peanut and corn design problems did not demonstrate significant differences across the majority of the metrics, and the alarm and coconut problems showed indifference across the same metrics. Additionally, all four problems were tested for equivalency with respect to the peanut problem using the Mann-Whitney test for equivalency. In this testing, it was found that none of the problems showed group mean equivalence, even after linear equating. This nonequivalence appears to be driven primarily from distribution differences among the metrics, which may be alleviated with larger sample sizes.

To understand why some problems exhibit similarities in metrics while others do not, a survey was included in the between-subjects testing to relate the responses to some of the design characteristics: problem difficulty, familiarity with the problem, and familiarity with existing solutions. Due to lack of statistical difference for the peanut and corn and the coconut and alarm problems in metric scores, the characteristics pf these problem pairings were investigated in greater detail. The peanut and corn problems were not statistically different on the difficulty of the problem and the familiarity of the participant with existing solutions. The alarm and coconut problems were found to be statistically different on all three characteristics, however. This indicates that while these characteristics may have an impact on the scoring, the relationship is inherently more complex than a direct relationship or simple interaction between the three characteristics observed.

This study was able to show equivalence between two problems tested, the peanut and corn problems, in the quantity and number of solutions metrics. These correlations are the first steps towards equivalent problems and with minor adjustments, the problems may be made equivalent on more metrics. Additionally, this study lays the framework for testing future design problems for equivalency. Within-subject analysis should serve as the foundation of equivalency testing in future research, using correlations such as Pearson and Spearman since it demonstrates that a given participant will score in a similar manner on both design problems. Participant characteristics' interactions with the design problems also warrant further exploration. Participant characteristics could include major, experience, familiarity with the design problem and solutions, and cultural background. Between-subject testing should not be used to assess equivalency but may

be utilized to prescreen problems as candidates for further equivalency testing. Future equivalence testing should also include characteristic testing to help researchers understand reasons for equality or inequality.

CHAPTER 6. FUTURE WORK

While this study presents two design problems as equivalent, more work is still needed to build a more complete set of equivalent problems. This work should be done in three steps; namely, problem characterization should be expanded to test for a greater number of characteristics, within-subjects testing should be pursued with more problems, and the ideation metrics should be refined.

By expanding the characterization of the design problems, the relationship between characteristics and metric scores can be fleshed out more completely. This would allow researchers to assess the complex relationships between characteristics and metric scores that were inaccessible in this study. Additionally, this would illustrate if certain characteristics that were not tested in this study, such as cultural background and expertise, have a large influence on the results. By understanding these relationships, researchers may more easily be able to make adjustments leading to more equivalent design problems.

The within-subjects testing conducted in this study gave valuable insight into the relationships between the corn and peanut problems.. An expansion of the within-subjects testing to include different problem pairings as well as problem characteristics would greatly benefit the field. I recommend the alarm and coconut problems to be examined next in within-subject analysis. As these problems showed similar statistical indifference as the peanut and corn problems, their results at the within-subject level can assess the viability of between-subjects testing as an indicator for within-subjects results. Additionally, by expanding the research on problem characteristics to within subjects

testing, changes in metric scores can be better correlated to changes in characteristic scores. For example, a particular problem characteristic, such as difficulty may be the same for two problems on average. However, within-subjects testing will expose if differences in perceived difficulty for an individual participant can be attributed to different metric scores. This form of tracking is unavailable using only between-subjects analysis.

To increase the reliability of studies like this, the ideation metrics require further refinements to ensure high inter-rater agreement. Before it can be fully accepted that two problems are similar, the design community must first be able to grade responses to individual problems in a consistent manner. This would bring about more repeatable studies and allow researchers to use previously graded responses without having to worry about variations brought about by the grader themselves. One way to improve upon inter-rater reliability is with the construction of test norms for each design problem. Test norms are sets of scores obtained from sample data with the procedure for obtaining these scores documented [79]. Once a set of design problems, such as the four in this study, are chosen by the community, norms can be established through careful collaborative assessments of solutions put forth. Through the establishment of good norms, any rater can more readily grade ideation without previous exposure.

In addition to these three areas of extended research, it is the belief of the author that a repetition of this study with even larger sample sizes would be of great value to the field. As was made apparent in equivalence testing, distribution inconsistencies are a major obstacle to overcome when making equivalent problems. If the study could be run

on sample sizes large enough to tend to normal distributions, it is the belief of the author that statements of equivalency could be made with greater conviction.

APPENDIX A. ORIGINAL DESIGN PROBLEMS

A.1 Original Peanut Design Problem

Design Problem - Device to Shell Peanuts

Problem Description:
In places like Haiti and certain West African countries, peanuts are a significant crop. Most peanut farmers shell their peanuts by hand, an inefficient and labor-intensive process. The goal of this project is to design and build a low-cost, easy to manufacture peanut shelling machine that will increase the productivity of the African peanut farmers. The target throughput is approximately 50 kg (110 lbs) per hour.

Customer Needs:

- Must remove the shell with minimal damage to the peanuts
- Electrical outlets are not available as a power source.
- A large quantity of peanuts must be quickly shelled.
- Low cost.
- Easy to manufacture

A.2 Original Corn Design Problem

Design Problem - Device to Aid in Shucking Corn

Problem Description:
Corn is currently the most widely grown crop in the Americas with the United States producing 40% of the world's harvest. However, only the loose corn kernels are used when bought canned or frozen in grocery stores. An ear of corn has a protective outer covering of leaves, known as the husk, and strands of corn silk threads run between the husk and the kernels. The removal of husk and silk to clean the corn is known as shucking corn. Design a device that quickly and cheaply shucks corn for mass production.

http://www.art-photograph-gallery.com/pictures-of-corn.html

Customer Needs:

- Must remove husk and silk from corn cob with minimal damage to kernels.
- A large quantity of corn must be shucked quickly.
- Low cost.

Please sketch and note (with words) one design solution per page.

A.3 Original Alarm Design Problem

Design Problem – Personal Alarm Clock

Problem Description:
Alarm clocks are widely used to help individuals wake from slumber. However, when used in shared spaces like dorm rooms, they will often disturb those around them. The goal of this problem is to design a low-cost alarm clock for individual use that will not disturb others. The clock should be portable for use in a variety of situations such as on the bus, in the library, or in a classroom.

Customer Needs:

- Must wake up individual with no disturbance to others.
- Must be portable and lightweight.
- Must be safe for user.
- Electrical outlets are not available as a constant power source.
- Low cost.

Please sketch and note (with words) one design solution per page starting on the next page.

A.4 Original Coconut Design Problem

Design Problem - Device to Aid in Coconut Harvesting

Problem Description:
In certain places like the Philippines, Indonesia, and India, coconut harvesting is a major practice. The current process requires a skilled person to climb the tree and cut down the coconuts. The average height of a coconut tree is 35-40 feet and though there are grooves along the tree that make it easier to climb, the tree surface becomes very slippery during the rainy seasons. The process may take as long as 12 hours for large farms that average 150 trees. The goal of this problem is to design a low-cost product to improve the coconut harvesting process so that it is safer and can be done more quickly. The target throughput is at least 500 pounds per hour.

Customer Needs:

- Must climb tree and remove coconut with little damage to fruit.
- Electrical outlets are not available as a power source.
- Low cost.

Please sketch and note (with words) one design solution per page.

CHAPTER 7. APPENDIX B. NEW DESIGN PROBLEMS

B.1 New Peanut Design Problem

Design Problem - Device to Shell Peanuts

Problem Description:
In places like Haiti and certain West African countries, peanuts are a significant crop. Most peanut farmers shell their peanuts by hand, an inefficient and labor-intensive process. The goal of this project is to design and build a low-cost, easy to manufacture peanut shelling machine that will increase the productivity of the African peanut farmers. The target throughput is approximately 50 kg (110 lbs) per hour.

Customer Needs:

- Must remove the shell with minimal damage to the peanuts.
- Electrical outlets are not available as a power source.
- A large quantity of peanuts must be quickly shelled.
- Low cost.
- Easy to manufacture.

Please sketch and note (with words) one design solution per page starting on the next page.

B.2 New Corn Design Problem

Design Problem - Device to Aid in Shucking Corn

Problem Description:
Corn is currently the most widely grown crop in the Americas with the United States producing 40% of the world's harvest. An ear of corn has a protective outer covering of leaves, known as the husk, and strands of corn silk threads run between the husk and the kernels. The removal of husk and silk to clean the corn is known as shucking corn. Design a device that quickly and cheaply shucks corn for mass production.

http://www.art-photograph-gallery.com/pictures-of-corn.html

Customer Needs:

- Must remove husk and silk from corn cob with minimal damage to kernels.
- A large quantity of corn must be shucked quickly.
- Must be safe for user
- Easy to manufacture
- Low cost

Please sketch and note (with words) one design solution per page starting on the next page.

B.3 New Alarm Design Problem

Design Problem – Personal Alarm Clock

Problem Description:
Alarm clocks are widely used to help individuals wake from slumber. However, when used in shared spaces like dorm rooms, they will often disturb those around them. The goal of this problem is to design a low-cost alarm clock for individual use that will not disturb others. The clock should be portable for use in a variety of situations such as on the bus, in the library, or in a classroom.

Customer Needs:

- Must wake up individual with no disturbance to others.
- Must be portable and lightweight.
- Must be safe for user.
- Electrical outlets are not available as a constant power source.
- Low cost

Please sketch and note (with words) one design solution per page starting on the next page.

B.4 New Coconut Design Problem

Design Problem - Device to Aid in Coconut Harvesting

Problem Description:
In certain places like the Philippines, Indonesia, and India, coconut harvesting is a major practice. The current process requires a skilled person to climb the tree and cut down the coconuts. The average height of a coconut tree is 35-40 feet and though there are grooves along the tree that make it easier to climb, the tree surface becomes very slippery during the rainy seasons. The current process may take as long as 12 hours for large farms that average 150 trees. The goal of this problem is to design a low-cost product to improve the coconut harvesting process so that it is safer and can be done more quickly.

Customer Needs:

- Must remove coconut with little damage to fruit and tree
- Must be safer to operate than current method.
- Must harvest coconuts quicker than current method
- Electrical outlets are not available as a power source.
- Low cost

Please sketch and note (with words) one design solution per page starting on the next page.

APPENDIX C SURVEY INSTRUMENTS

Accompanying Survey

Name: _____

GTID: _____

1.) Did you hear about this design problem ahead of time?

YES NO

2.) If yes, did you generate solutions before the session?

YES NO

3.) How would you rate the difficulty of this problem in terms of generating feasible solutions?

Very Easy	Easy	Neutral	Difficult	Very Difficult

4.) How would you rate your familiarity with the design problem?

Not at all familiar	Slightly Familiar	Somewhat Familiar	Moderately Familiar	Extremely Familiar

5.) How would you rate your familiarity with existing solutions?

Not at all familiar	Slightly Familiar	Somewhat Familiar	Moderately Familiar	Extremely Familiar

Comments:

APPENDIX D BIN LISTS

D.1 Peanut Bin List

Adhesive	Mass difference
Black box	Metal screen/grate/force through
Blade	Mixer (eggbeater)
Boil to remove shell	Needles/hooks/spikes
Brittle shell	No shell peanuts
Buoyancy (in liquid)	Press/weight
Burn shells	Pressurized fluid
Centrifuge	Punch
Chemical	Reframing the problem
Conical rollers	Robot hands
Cylindrical or spherical rollers	Scrape/brush
Eating/decomposition	Shear
Filter/mesh	Softening shell
Flexible tube	Spring around shell
Force on ends	Squeeze from one end
Friction/abrasive	Thermal expansion
High frequency radio waves	Toothed rollers/gears
High velocity impact	Torsion
Hinged plates	Train animal
Hold shell in place, nut falls out	Tumblers
Human	Vacuum
Laser	Vibration/shake
Magnet	

D.2 Corn Bin List

Abrasives	Manual clamp
Adhesive	Moving fluid (air & water)
Animal shucking	Perforate
Automatic size adjustment	Popcorn
Automatic solution/black box	Remove core
Blade	Restructure problem
Break stem	Robotic hands
Brushes	Rollers
Chemicals	High frequency waves
Drill Flutes	High-Speed Rotation/Centrifuge
Eating	Scraper
Electricity	Series of mesh wires
Embrittle husk	Slit plates
Entanglement	Soften husk
Filter	Sorter
Fire	Teleport
Impact	Torsion
Humans	Tumbler
Huskless/Modified husk	Vibration
Lasers	Visual sensor
Magnets	

D.3 Alarm Bin List

Adjust Volume/proximity	Projectile
Alter Dreams	REM Cycle
Biological Clock	Restrict Breathing
Change Posture	Scents
Constricting Band	Sensitizing Drug
Dierrhetic	Sleeping Aid
Directed Sound	Spotlight
Electrocution	Stabbing/Studs
Falling Sensation	Standalone Vibrator
Flashing Contacts	Sunlight
Flavor Capsule	Temperature
Hair Pulling	Tickling
Headphones	Trained Animal
Human	Wakeup Drug
Impact	Water Sprayer
Increase Breathing	Wearable Lights (glasses, bracelet, etc.)
Insects	Wearable Sprayer
Nanobots	Wearable Vibrator
Noise Cancellation	Winds
Personalized Frequency	

D.4 Coconut Bin List

Adhesive (gloves, pants, shoes)	Pressurized fluid
Basket/backpack	RC helicopter or plane
Blade (device)	Reframing the problem
Bury tree in dirt	Robot
Cable lift/suspension	Shoes with springs
Climbing equipment on tree	Shoes, belts, gloves with spikes/grabbers
Gun - attach to coconut	Slide/chute
Gun - impact	Spikes
Handheld blade	Stilts
Hot air balloon	Thrown object
Impact	Tools on a pole (blade, saw)
Jet pack	Tools on a pole (grabber)
Ladder/Stair	Trained animal
Laser	Tree Climber
Lift	Tree shaker
Modify trees or coconuts	Vacuum
Movable enclosure	Vehicle
Net/canopy/inflatable pad	Trampoline
Platform/walkway system	

APPENDIX E ORIGINAL INTER-RATER RELIABILITY

Table 16. Original Inter-rater Reliabilty

	Between-Subjects								Within-Subjects			
	Peanut	n	Corn	n	Alarm	n	Coconut	n	Peanut	n	Corn	n
Quantity	0.75	10	0.82	10	.93	10	.94	10	0.82	10	0.84	10
Quality	0.18	10	-0.03	10	.62	10	0.22	10	0.44	10	0.23	10
Novelty	0.81	10	0.74	10	.91	10	.65	10	0.78	10	0.92	10
Variety	0.58	10	0.64	10	.99	10	.80	10	0.96	10	0.62	10
Percent of Sample	20.00		20.83/100		21.74		21.74		25.64		25.64	

Inter-rater reliability was originally checked using by having a second grader rate the solutions of ten participants for all metrics and problems. The initial ratings for corn quality in the between-subject data showed no correlation and as a result, the second grader was asked to grade all participants in this metric. Table 16 shows the initial results and Table 2 shows the updated values. By increasing the percent of the sample graded, the inter-rater reliability was greatly increased.

REFERENCES

[1] B. Mullen, C. Johnson, and E. Salas, "Productivity loss in brainstorming groups: A meta-analytic integration," *Basic and applied social psychology,* vol. 12, pp. 3-23, 1991.

[2] G. Scott, L. E. Leritz, and M. D. Mumford, "The effectiveness of creativity training: A quantitative review," *Creativity Research Journal,* vol. 16, pp. 361-388, 2004/12/01 2004.

[3] F. Durand, M. E. Helms, J. Tsenn, D. A. McAdams, and J. S. Linsey, "In Search of Effective Design Problems for Design Research," in *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2015, p. V007T06A011.

[4] O. Atilola, M. Tomko, and J. S. Linsey, "The effects of representation on idea generation and design fixation: A study comparing sketches and function trees," *Design Studies,* vol. 42, pp. 110-136, 2016.

[5] D. G. Jansson and S. M. Smith, "Design fixation," *Design studies,* vol. 12, pp. 3-11, 1991.

[6] M. K. Perttula and L. A. Liikkanen, "Structural tendencies and exposure effects in design idea generation," in *ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2006, pp. 199-210.

[7] J. D. Summers and J. J. Shah, "Mechanical engineering design complexity metrics: size, coupling, and solvability," *Journal of Mechanical Design,* vol. 132, p. 021004, 2010.

[8] J. O. Wilson, D. Rosen, B. A. Nelson, and J. Yen, "The effects of biological examples in idea generation," *Design Studies,* vol. 31, pp. 169-186, 2010.

[9] N. Cross, "Expertise in design: an overview," *Design studies,* vol. 25, pp. 427-441, 2004.

[10] V. Viswanathan and J. Linsey, "A study on the role of expertise in design fixation and its mitigation," in *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2012, pp. 901-911.

[11] R. Morocz, B. D. Levy, C. R. Forest, R. L. Nagel, W. C. Newstetter, K. G. Talley, *et al.*, "University Maker Spaces: Discovery, Optimization and Measurement of Impacts," in *2015 ASEE Annual Conference & Exposition*, Seattle, Washington, 2015.

[12] J. J. Shah, "Experimental Investigation of Progressive Idea Generation Techniques in Engineering Design," presented at the DETC'98, 1998 ASME Design Engineering Technical Conferences, Atlanta, GA, 1998.

[13] J. J. Shah, N. Vargas-Hernández, J. S. Summers, and S. Kulkarni, "Collaborative Sketching (C-Sketch) – An Idea Generation Technique for Engineering Design," *Journal of Creative Behavior*, vol. 35, pp. 168-198, 2001.

[14] L. A. Vasconcelos and N. Crilly, "Inspiration and fixation: Questions, methods, findings, and challenges," *Design Studies*, vol. 42, pp. 1-32, 2016.

[15] O. Atilola and J. Linsey, "Representing Analogies to Influence Fixation and Creativity: A Study Comparing Computer-Aided Design, Photographs, and Sketches," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, vol. 29, pp. 161-171, 2015.

[16] R. Hannah, S. Joshi, and J. D. Summers, "A user study of interpretability of engineering design representations," *Journal of Engineering Design*, vol. 23, pp. 443-468, 2012.

[17] S. Joshi and J. D. Summers, "Impact of Requirements Elicitation Activity on Idea Generation: A Designer Study," in *ASME 2014 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Buffalo, New York, 2014, p. V007T07A026.

[18] U. N. Sio, K. Kotovsky, and J. Cagan, "Fixation or inspiration? A meta-analytic review of the role of examples on design processes," *Design Studies*, vol. 39, pp. 70-99, 2015.

[19] V. Kumar and G. Mocko, "Similarity of Engineering Design Problems to Enable Reuse in Design Research Experiments," in *ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Charlotte, North Carolina, 2016, p. V007T06A042.

[20] J. J. Shah, S. V. Kulkarni, and N. Vargas-Hernandez, "Evaluation of idea generation methods for conceptual design: effectiveness metrics and design of experiments," *Journal of mechanical design*, vol. 122, pp. 377-384, 2000.

[21] J. J. Shah, S. M. Smith, and N. Vargas-Hernandez, "Metrics for measuring ideation effectiveness," *Design studies*, vol. 24, pp. 111-134, 2003.

[22] C. J. Atman, J. R. Chimka, K. M. Bursic, and H. L. Nachtmann, "A comparison of freshman and senior engineering design processes," *Design studies,* vol. 20, pp. 131-152, 1999.

[23] J. Tsenn, O. Atilola, D. A. McAdams, and J. S. Linsey, "The effects of time and incubation on design concept generation," *Design Studies,* vol. 35, pp. 500-526, 2014.

[24] V. Viswanathan, O. Atilola, N. Esposito, and J. Linsey, "A study on the role of physical models in the mitigation of design fixation," *Journal of Engineering Design,* pp. 1-19, 2014.

[25] V. K. Viswanathan and J. S. Linsey, "Design fixation and its mitigation: a study on the role of expertise," *Journal of Mechanical Design,* vol. 135, p. 051008, 2013.

[26] N. Genco, K. Hölttä-Otto, and C. C. Seepersad, "An Experimental Investigation of the Innovation Capabilities of Undergraduate Engineering Students," *Journal of Engineering Education,* vol. 101, pp. 60-81, 2012.

[27] A. Erlebacher, "Design and analysis of experiments contrasting the within-and between-subjects manipulation of the independent variable," *Psychological Bulletin,* vol. 84, p. 212, 1977.

[28] A. G. Greenwald, "Within-subjects designs: To use or not to use?," *Psychological Bulletin,* vol. 83, p. 314, 1976.

[29] T. B. Sprecher, "A study of engineers' criteria for creativity," *Journal of Applied Psychology,* vol. 43, p. 141, 1959.

[30] D. H. Cropley, "Creativity in engineering," in *Multidisciplinary Contributions to the Science of Creative Thinking*, ed: Springer, 2016, pp. 155-173.

[31] G. Pahl and W. Beitz, *Engineering design: a systematic approach*: Springer Science & Business Media, 2013.

[32] M. Baker. First results from psychology's largest reproducibility test. *Nature*.

[33] A. Newell and H. A. Simon, *Human problem solving* vol. 104: Prentice-Hall Englewood Cliffs, NJ, 1972.

[34] H. A. Simon, "The structure of ill structured problems," *Artificial intelligence,* vol. 4, pp. 181-201, 1973.

[35] D. A. Schon, *The reflective practitioner: How professionals think in action* vol. 5126: Basic books, 1984.

[36] K. Dorst and N. Cross, "Creativity in the design process: co-evolution of problem–solution," *Design studies,* vol. 22, pp. 425-437, 2001.

[37] M. L. Maher, J. Poon, and S. Boulanger, "Formalising design exploration as co-evolution," in *Advances in formal design methods for CAD,* ed: Springer, 1996, pp. 3-30.

[38] K. Dorst, "The problem of design problems," *Expertise in design,* pp. 135-147, 2003.

[39] J. L. Mathieson, B. A. Wallace, and J. D. Summers, "Assembly time modeling through connective complexity metrics," in *Manufacturing Automation (ICMA), 2010 International Conference on*, 2010, pp. 16-23.

[40] F. Ameri, J. D. Summers, G. M. Mocko, and M. Porter, "Engineering design complexity: an investigation of methods and measures," *Research in Engineering Design,* vol. 19, pp. 161-179, 2008.

[41] J. D. Summers and J. J. Shah, "Developing measures of complexity for engineering design," in *ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2003, pp. 381-392.

[42] J. Mathieson, M. Miller, and J. Summers, "A Protocol for Connective Complexity Tracking in the Engineering Design Process," in *DS 68-7: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 7: Human Behaviour in Design, Lyngby/Copenhagen, Denmark*, 2011.

[43] D. H. Jonassen, "Toward a design theory of problem solving," *Educational technology research and development,* vol. 48, pp. 63-85, 2000.

[44] R. E. Mayer, *Thinking, problem solving, cognition*: WH Freeman/Times Books/Henry Holt & Co, 1992.

[45] M. U. Smith, *Toward a unified theory of problem solving: Views from the content domains*: Routledge, 2012.

[46] M. U. Smith, "Toward a Unified Theory of Problem Solving: A View from Biology," 1988.

[47] J. Funke, "Solving complex problems: Exploration and control of complex systems," *Complex problem solving: Principles and mechanisms,* pp. 185-222, 1991.

[48] C. A. Toh, A. A. Strohmetz, and S. R. Miller, "The Effects of Gender and Idea Goodness on Ownership Bias in Engineering Design Education," *Journal of Mechanical Design,* vol. 138, pp. 101105-101105-8, 2016.

[49] D. G. Johnson, N. Genco, M. N. Saunders, P. Williams, C. C. Seepersad, and K. Hölttä-Otto, "An Experimental Investigation of the Effectiveness of Empathic Experience Design for Innovative Concept Generation," *Journal of Mechanical Design*, vol. 136, pp. 051009-051009-12, 2014.

[50] L. A. Liikkanen, T. A. Björklund, M. M. Hämäläinen, and M. P. Koskinen, "Time constraints in design idea generation," in *DS 58-9: Proceedings of ICED 09, the 17th International Conference on Engineering Design, Vol. 9, Human Behavior in Design*, Palo Alto, CA, 2009.

[51] M. K. Perttula and L. A. Liikkanen, "Exposure effects in design idea generation: Unconscious conformity or a product of sampling probability?," *Development Process: From Idea to the World's First Bionic Prosthetic Foot,* 2006.

[52] P. Howard-Jones and S. Murray, "Ideational productivity, focus of attention, and context," *Creativity research journal*, vol. 15, pp. 153-166, 2003.

[53] J. Kelly, G. C. Futoran, and J. E. McGrath, "Capacity and capability seven studies of entrainment of task performance rates," *Small Group Research,* vol. 21, pp. 283-314, 1990.

[54] A. Snyder, J. Mitchell, S. Ellwood, A. Yates, and G. Pallier, "Nonconscious idea generation," *Psychological reports,* vol. 94, pp. 1325-1330, 2004.

[55] J. P. Guilford, "Some incubated thoughts on incubation," *The Journal of Creative Behavior,* vol. 13, pp. 1-8, 1979.

[56] V. K. Viswanathan and J. S. Linsey, "Physical models and design thinking: A study of functionality, novelty and variety of ideas," *Journal of Mechanical Design*, vol. 134, p. 091004, 2012.

[57] S. M. Smith and S. E. Blankenship, "Incubation and the persistence of fixation in problem solving," *The American journal of psychology*, pp. 61-87, 1991.

[58] N. Kohn and S. M. Smith, "Partly versus completely out of your mind: Effects of incubation and distraction on resolving fixation," *The Journal of Creative Behavior,* vol. 43, pp. 102-118, 2009.

[59] A. T. Purcell and J. S. Gero, "Effects of Examples on the Results of a Design Activity," *Knowledge-Based Systems,* vol. 5, pp. 82-91, 1992.

[60] A. T. Purcell and J. S. Gero, "Design and other types of fixation," *Design studies*, vol. 17, pp. 363-383, 1996.

[61] M. Perttula and P. Sipilä, "The idea exposure paradigm in design idea generation," *Journal of Engineering Design,* vol. 18, pp. 93-102, 2007.

[62] K. Fu, J. Cagan, and K. Kotovsky, "Design team convergence: the influence of example solution quality," *Journal of Mechanical Design,* vol. 132, p. 111005, 2010.

[63] J. S. Linsey, I. Tseng, K. Fu, J. Cagan, K. L. Wood, and C. Schunn, "A study of design fixation, its mitigation and perception in engineering design faculty," *Journal of Mechanical Design,* vol. 132, p. 041003, 2010.

[64] K. L. Dugosh and P. B. Paulus, "Cognitive and social comparison processes in brainstorming," *Journal of experimental social psychology,* vol. 41, pp. 313-320, 2005.

[65] K. L. Dugosh, P. B. Paulus, E. J. Roland, and H.-C. Yang, "Cognitive stimulation in brainstorming," *Journal of personality and social psychology,* vol. 79, p. 722, 2000.

[66] F. L. McKoy, N. Vargas-Hernández, J. D. Summers, and J. J. Shah, "Influence of design representation on effectiveness of idea generation," in *ASME IDETC Design Theory and Methodology Conference, Pittsburgh, PA, Sept*, 2001, pp. 9-12.

[67] C. Cardoso and P. Badke-Schaub, "The influence of different pictorial representations during idea generation," *J. Creat. Behav,* vol. 45, pp. 130-146, 2011.

[68] G. Goldschmidt and M. Smolkov, "Variances in the impact of visual stimuli on design problem solving performance," *Design Studies,* vol. 27, pp. 549-569, 2006.

[69] J. Linsey, J. Murphy, A. Markman, K. Wood, and T. Kurtoglu, "Representing analogies: Increasing the probability of innovation," in *ASME 2006 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2006, pp. 265-282.

[70] J. Linsey, K. L. Wood, and A. B. Markman, "Modality and representation in analogy," *Artificial Intelligence for Engineering Design, Analysis and Manufacturing,* vol. 22, pp. 85-100, 2008.

[71] K. L. Wood, "Development of a functional basis for design," *Journal of Mechanical design,* vol. 122, pp. 359-370, 2000.

[72] R. Srivathsavai, N. Genco, K. Hölttä-Otto, and C. C. Seepersad, "Study of existing metrics used in measurement of ideation effectiveness," in *ASME 2010 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2010, pp. 355-366.

[73] H. H. Friedman and T. Amoo, "Rating the rating scales," *Journal of Marketing Management,* pp. 114-123, 1999.

[74] B. A. Nelson, J. O. Wilson, D. Rosen, and J. Yen, "Refined metrics for measuring ideation effectiveness," *Design Studies,* vol. 30, pp. 737-743, 2009.

[75] J. Peeters, P.-A. Verhaegen, D. Vandevenne, and J. Duflou, "Refined metrics for measuring novelty in ideation," in *IDMME Virtual Concept 2010*, Bordeaux, France, 2010.

[76] C. Charyton, R. J. Jagacinski, and J. A. Merrill, "CEDA: A research instrument for creative engineering design assessment," *Psychology of Aesthetics, Creativity, and the Arts,* vol. 2, p. 147, 2008.

[77] C. Charyton and J. A. Merrill, "Assessing general creativity and creative engineering design in first year engineering students," *Journal of engineering education,* vol. 98, pp. 145-156, 2009.

[78] D. C. Brown, "Problems with the calculation of Novelty metrics," in *Proc. Design Creativity Workshop, 6th Int. Conf. on Design Computing and Cognition (DCC'14),* 2014.

[79] P. Kline, *A handbook of test constructio : introduction to psychometric design*: Methuan & Co, 1986.

[80] A. Anastasi, "Evolving concepts of test validation," *Annual review of Psychology,* vol. 37, pp. 1-16, 1986.

[81] S. Wellek, "A New Approach to Equivalence Assessment in Standard Comparative Bioavailability Trials by Means of the Mann-Whitney Statistic," *Biometrical journal,* vol. 38, pp. 695-710, 1996.

[82] S. Wellek, *Testing statistical hypotheses of equivalence and noninferiority*, Second ed.: CRC Press, 2010.

[83] J. S. Linsey, M. G. Green, J. Murphy, K. L. Wood, and A. B. Markman, ""Collaborating To Success": An Experimental Study of Group Idea Generation Techniques," in *ASME 2005 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2005, pp. 277-290.

[84] J. S. Linsey, K. Wood, and A. Markman, "Increasing innovation: presentation and evaluation of the wordtree design-by-analogy method," in *ASME 2008 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2008, pp. 21-32.

[85] R. Lopez, J. S. Linsey, and S. M. Smith, "Characterizing the effect of domain distance in design-by-analogy," in *ASME 2011 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2011, pp. 141-151.

[86] O. Atilola, V. Viswanathan, and J. Linsey, "A study on the representation of examples in learning engineering concepts," in *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2012, pp. 37-46.

[87] F. Durand, M. E. Helms, J. Tsenn, E. McTigue, D. A. McAdams, and J. S. Linsey, "Teaching Students to Innovate: Evaluating Methods for Bioinspired Design and Their Impact on Design Self Efficacy," in *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2015, p. V007T06A003.

[88] M. W. Glier, D. A. McAdams, and J. S. Linsey, "An experimental investigation of analogy formation using the engineering-to-biology thesaurus," in *ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2013, p. V005T06A007.

[89] M. W. Glier, J. Tsenn, J. S. Linsey, and D. A. McAdams, "Evaluating the directed method for bioinspired design," in *ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2012, pp. 403-413.

[90] M. W. Glier, J. Tsenn, D. A. McAdams, and J. S. Linsey, "Evaluating methods for bioinspired concept generation," in *Design Computing and Cognition'12*, ed: Springer, 2014, pp. 41-57.

[91] J. W. Kim, D. A. McAdams, and J. Linsey, "Helping students to find biological inspiration: Impact of valuableness and presentation format," in *Frontiers in Education Conference (FIE), 2014 IEEE*, 2014, pp. 1-6.

[92] N. Genco, D. Johnson, K. Hölttä-Otto, and C. C. Seepersad, "A study of the effectiveness of empathic experience design as a creativity technique," in *ASME 2011 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, 2011, pp. 131-139.

[93] M. W. Glier, J. Tsenn, J. S. Linsey, and D. A. McAdams, "Methods for supporting bioinspired design," in *ASME 2011 International Mechanical Engineering Congress and Exposition*, 2011, pp. 737-744.

[94] W. M. Vagias, "Likert-type Scale Response Anchors. Clemson International Institute for Tourism," *& Research Development, Department of Parks, Recreation and Tourism Management, Clemson University*, 2006.

[95] *The Engineer of 2020: Visions of Engineering in the New Century*: The National Academies Press, 2004.

[96] J. S. Linsey, E. Clauss, T. Kurtoglu, J. Murphy, K. Wood, and A. Markman, "An experimental study of group idea generation techniques: understanding the roles

of idea representation and viewing methods," *Journal of Mechanical Design,* vol. 133, p. 031008, 2011.

[97] R. Pucha, B. D. Levy, J. Linsey, S. Newton, M. Alemdar, and T. Utschig, "Assessing Concept Generation Intervention Strategies for Creativity Using Design Problems in Freshman Engineering Graphics Course," in *2017 ASEE Annual Conference & Exposition*, Columbus, OH, 2017.

[98] S. R. Daly, J. L. Christian, S. Yilmaz, C. M. Seifert, and R. Gonzalez, "Assessing design heuristics for idea generation in an introductory engineering course," *International Journal of Engineering Education,* vol. 28, 2012.

[99] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Second ed. Hillsdale, New Jersey 07642: Lawrence Erlbaum Associates, Inc., Publishers, 1988.